Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users?

El Boghdady, Nawal; Gaudrain, Etienne; Baskent, Deniz

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/1.5087693

Publication date:
2019

Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users?

Nawal El Boghdady, Etienne Gaudrain, and Deniz Başkent

Citation: The Journal of the Acoustical Society of America **145**, 417 (2019); doi: 10.1121/1.5087693
View online: https://doi.org/10.1121/1.5087693
View Table of Contents: https://asa.scitation.org/toc/jas/145/1
Published by the Acoustical Society of America

Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users?[a)]

Nawal El Boghdady,[b),c)] Etienne Gaudrain,[b),d)] and Deniz Başkent[b)]
Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

(Received 3 August 2018; revised 20 December 2018; accepted 21 December 2018; published online 25 January 2019)

Differences in voice pitch (*F*0) and vocal tract length (VTL) improve intelligibility of speech masked by a background talker (speech-on-speech; SoS) for normal-hearing (NH) listeners. Cochlear implant (CI) users, who are less sensitive to these two voice cues compared to NH listeners, experience difficulties in SoS perception. Three research questions were addressed: (1) whether increasing the *F*0 and VTL difference (Δ*F*0; ΔVTL) between two competing talkers benefits CI users in SoS intelligibility and comprehension, (2) whether this benefit is related to their *F*0 and VTL sensitivity, and (3) whether their overall SoS intelligibility and comprehension are related to their *F*0 and VTL sensitivity. Results showed: (1) CI users did not benefit in SoS perception from increasing Δ*F*0 and ΔVTL; increasing ΔVTL had a slightly detrimental effect on SoS intelligibility and comprehension. Results also showed: (2) the effect from increasing Δ*F*0 on SoS intelligibility was correlated with *F*0 sensitivity, while the effect from increasing ΔVTL on SoS comprehension was correlated with VTL sensitivity. Finally, (3) the sensitivity to both *F*0 and VTL, and not only one of them, was found to be correlated with overall SoS performance, elucidating important aspects of voice perception that should be optimized through future coding strategies.

https://doi.org/10.1121/1.5087693

[VB] Pages: 417–439

I. INTRODUCTION

Cochlear implant (CI) users have more difficulties understanding speech in multi-talker settings compared to normal hearing (NH) listeners (e.g., Cullington and Zeng, 2008; Stickney *et al.*, 2004; Stickney *et al.*, 2007), yet the relationship between this difficulty and voice cue perception remains relatively unknown. In normal hearing (NH), for such speech-on-speech (SoS) perception, the voice cues related to the target (foreground) and masker (interfering) speakers seem to play an important role. This was demonstrated by higher SoS intelligibility when the voices of each of the target and masker belonged to different speakers, especially if they were of the opposite gender[1] (Brungart, 2001; Brungart *et al.*, 2009; Festen and Plomp, 1990; Stickney *et al.*, 2004).

Among many voice characteristics that help define/identify a voice (Abercrombie, 1967) and can be used for a benefit in SoS perception, two fundamental voice characteristics seem to be most important. The first voice characteristic is the speaker's fundamental frequency (*F*0), which gives cues to the voice pitch. The second voice characteristic is the speaker's vocal tract length (VTL), which is associated with the physical (Fitch and Giedd, 1999) and perceived size of a speaker (Ives *et al.*, 2005; Smith *et al.*, 2005). *F*0 cues are represented in both the temporal envelope of the signal and the corresponding place of stimulation along the cochlea (e.g., Carlyon and Shackleton, 1994; Licklider, 1954; Oxenham, 2008), while VTL cues are mainly encoded in the relationship between the formant peaks in the spectral envelope of the signal (Chiba and Kajiyama, 1941; Fant, 1960; Lieberman and Blumstein, 1988; Müller, 1848; Stevens and House, 1955). Because the representation of *F*0 in the speech signal is different from that of VTL, their perceptual effects can also be expected to differ.

*F*0 and VTL cues have been found to contribute to talker gender categorization in NH listeners (Fuller *et al.*, 2014; Hillenbrand and Clark, 2009; Meister *et al.*, 2016; Skuk and Schweinberger, 2014; Smith *et al.*, 2007; Smith and Patterson, 2005). Moreover, when differences in either of these two voice cues are introduced between target and masker speakers in SoS tasks, NH listeners demonstrate an increase in target sentence identification scores, supporting the importance of these voice cues in SoS perception (e.g., Başkent and Gaudrain, 2016; Brokx and Nooteboom, 1982; Darwin *et al.*, 2003; Drullman and Bronkhorst, 2004; Vestergaard *et al.*, 2009).

Speech delivered via electric stimulation of a CI is inherently degraded in spectrotemporal resolution (for a review, see Başkent *et al.*, 2016), which is expected to affect

[a)]Portions of the results of this study were presented in "On the colour of voices: Does good perception of vocal differences relate to better speech intelligibility in cocktail-party settings?," 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan, 2016.
[b)]Also at: Graduate School of Medical Sciences, Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Groningen, The Netherlands.
[c)]Electronic mail: n.el.boghdady@umcg.nl
[d)]Also at: CNRS UMR 5292, INSERM U1028, Lyon Neuroscience Research Center, Université de Lyon, Lyon, France.

the perception of *F*0 and VTL differences and, correspondingly, their effective benefit in SoS perception. Directly supporting this idea, previous literature has shown that when stimuli were sufficiently degraded using acoustic vocoder simulations of CI processing, NH listeners became less sensitive to both *F*0 and VTL differences, compared to listening in the non-vocoded condition (Gaudrain and Başkent, 2015). In line with these findings, NH listeners exposed to vocoded SoS were also shown to benefit differently from voice cue differences between target and masker speakers, depending on the type of vocoder used. For example, sinewave vocoders, which were shown to partially preserve some of the spectrotemporal aspects of *F*0 cues (Gaudrain and Başkent, 2015), were also shown to preserve some benefit from talker differences between target and masker speakers (Cullington and Zeng, 2008). In contrast, noise-band vocoders, which do not preserve such voice cues (Gaudrain and Başkent, 2015), were also shown to contribute to the overall lack of benefit from either natural (Qin and Oxenham, 2003; Stickney *et al*., 2004) or synthesized (Qin and Oxenham, 2005; Stickney *et al*., 2007) voice cue differences between target and masker speakers.

Similar to what has been observed in the aforementioned vocoder studies, CI users, when compared to NH listeners, were also shown to not only have reduced sensitivity to *F*0 and VTL differences (Gaudrain and Başkent, 2018; Zaltz *et al*., 2018), but also impaired gender judgements based on these two cues (Fuller *et al*., 2014; Meister *et al*., 2016). Mixed results have been reported in CI users when voice cue differences were increased between target and masker speakers in SoS tasks (Cullington and Zeng, 2008; Pyschny *et al*., 2011; Stickney *et al*., 2004; Stickney *et al*., 2007). On the one hand, Cullington and Zeng (2008), who measured SoS intelligibility in a group of CI participants, reported a benefit in SoS intelligibility from changing the gender of the masker relative to that of the target. Similar findings for bimodal CI users listening with only their CI activated were also reported by Pyschny *et al*. (2011), who observed a benefit in SoS intelligibility as a function of increasing the masker's *F*0 relative to that of the target speaker. On the other hand, Stickney *et al*. reported no such benefit for CI users, either as a function of changing the gender of the masker relative to that of the target speaker (Stickney *et al*., 2004) or as a function of only changing the masker's *F*0 relative to that of the target (Stickney *et al*., 2007). One potential explanation for this discrepancy between studies may come from the differences in the CI samples tested. For example, Cullington and Zeng (2008) attributed the difference between their results and those of Stickney *et al*. (2004) and Stickney *et al*. (2007) to the slightly better performance of their CI participants in noise compared to that of the CI users recruited in either of the studies by Stickney *et al*. Moreover, the 12 CI participants tested by Pyschny *et al*. (2011) were all bimodal users, 8 of which had some useable residual acoustic hearing since their unaided thresholds were better than 90 dB hearing level (HL). Thus, it is possible that the benefit reported by Pyschny *et al*. is partly due to the participants' residual acoustic hearing rather than the CI processing *per se*.

However, in contrast to this reported benefit from *F*0 differences between target and masker, the same data from Pyschny *et al*. (2011) revealed a *decrement* in SoS intelligibility as a function of shortening the VTL of the masker relative to that of the target, both for the CI-only and bimodal conditions. These findings support the notion that the effects of *F*0 and VTL cues in SoS tasks may indeed be substantially different.

Nonetheless, Pyschny *et al*. (2011) had no NH control participants in their study and applied rather small VTL differences between target and masker speakers that are well below most CI users' typical VTL detection thresholds (Gaudrain and Başkent, 2018). Thus, the question remains whether the specific VTL manipulations by Pyschny *et al*. were expected to yield a benefit for NH listeners as well and whether CI listeners would gain an improvement in SoS intelligibility for larger VTL differences that encompass CI users' typical VTL detection thresholds.

CI users' typical *F*0 and VTL detection thresholds are around 9.19 semitones (st; one-twelfth of an octave) and 7.19 st, respectively (Gaudrain and Başkent, 2018). Based on the data of Peterson and Barney (1952), on the one hand, the maximum voice difference between a typical female and typical male is around 12 st for *F*0 and around 3.8 st for VTL. This means that while some CI users may be able to detect *F*0 differences between females and males, most of them might not be able to detect VTL differences. On the other hand, the maximum voice difference between a typical female and typical child is approximately 15 st for *F*0 and about 8.3 st for VTL, which means that, in principle, most CI users should be able to detect both *F*0 and VTL differences between females' and children's voices if these differences are large enough.

This study investigated the question of whether SoS perception is related to voice cue sensitivity in CI users. The hypothesis was that CI users' deficits in SoS intelligibility could relate to their reduced sensitivity in vocal cue perception. Three research questions were posed to test for the presence of this relationship.

The first question, addressed by experiments 1 and 2, was whether CI users would benefit from *F*0 and VTL differences ($\Delta F0$; ΔVTL) between target and masker speakers in SoS perception, in a similar manner to NH listeners. SoS performance was measured for both NH and CI listeners as a function of systematically increasing $\Delta F0$ and ΔVTL between target and masker speakers. The target and masker sentences were taken initially from the same speaker to overcome differences in speaking styles that may emerge from having different speakers (such as the speaking-rate difference mentioned by Cullington and Zeng, 2008). The range for *F*0 and VTL differences was chosen to encompass CI users' typical sensitivity thresholds reported in the literature (Gaudrain and Başkent, 2018; Zaltz *et al*., 2018). This range was chosen to ensure that the *F*0 and VTL differences introduced between target and masker voices would be detected by the CI users tested. Experiments 1 and 2 differed in speech materials and the specific task administered. This was carried out in an attempt to provide tasks that measure different aspects of speech perception, which may also

potentially differ in task difficulty, and hence improve the dynamic range of performance for observing effects in both groups. In experiment 1, SoS intelligibility was measured for NH and CI users in a manner similar to previous literature (Pyschny *et al.*, 2011; Stickney *et al.*, 2004; Stickney *et al.*, 2007). Participants were asked to repeat all of the words in the target sentence presented simultaneously with a single competing masker, and the intelligibility score was determined based on the number of words correctly repeated. In experiment 2, an alternative speech test was used, namely, a sentence verification task (SVT), which measures overall sentence comprehension (Adank and Janse, 2009; Baddeley *et al.*, 1992; May *et al.*, 2001; Pisoni *et al.*, 1987; Saxton *et al.*, 2001). In this task, participants were asked to judge whether the target sentence statement, presented simultaneously with a single competing masker, was true or false, without repeating the actual sentence, and both target sentence comprehension accuracy and speed (response times; RTs) were measured (e.g., as was done by Adank and Janse, 2009).

The second research question, addressed in experiment 3, was whether the effect of increasing *F*0 and VTL between target and masker on SoS perception (experiments 1 and 2) would correlate with CI users' sensitivity to *F*0 and VTL cues as measured by just-noticeable-difference (JND) measures. More specifically, participants with lower JNDs (i.e., more sensitive to *F*0 and VTL differences) would be more likely to benefit from *F*0 and VTL differences in SoS scenarios.

The final research question, also addressed in experiment 3, was whether the average overall SoS performance per participant across all voice conditions from experiments 1 and 2 would correlate with their *F*0 and VTL JNDs. The hypothesis was that higher sensitivity to *F*0 and VTL differences would correlate with higher SoS overall performance.

II. GENERAL METHODS

A. Participants

All NH and CI participants were native Dutch or Frisian speakers who used Dutch as the primary language of communication, and who had no reported health problems, such as dyslexia or attention deficit hyperactivity disorder.

1. NH listeners

NH control participants were recruited from the student body of the University of Groningen. Eighteen NH listeners (five males), aged 19 to 27 yr ($\mu = 22.67$ yr, $\sigma = 2.03$ yr), participated in experiments 1 and 2 only. NH participants had pure tone thresholds less than or equal to 20 dB HL at octave frequencies between 250 Hz and 8 kHz on either ear.

2. CI listeners

Participants with CIs were recruited both from the clinical database at the University Medical Center Groningen (UMCG) and the general public. This was done to ensure a better representation of the general CI population with a relatively large number of participants.

Participants were recruited based on their post-operative clinical speech perception scores in quiet, measured as the percentage of correctly repeated phonemes embedded in meaningful consonant-vowel-consonant (CVC) Dutch words from the Nederlandse Vereniging voor Audiologie (NVA) corpus (Bosman and Smoorenburg, 1995). The participants were selected to have a minimum NVA score of 40% (see Table I) to ensure that they could perform the experiments. In addition, a wide range of NVA scores was included to both have a more representative sample of CI participants and enough variability to test the correlation between the voice cue JNDs and SoS perception. Initially, the recruitment criteria included a minimum duration of device use of one year to ensure that the implantation outcome had mostly stabilized. However, this constraint was relaxed for participants with NVA scores that were higher than 60% to recruit a relatively larger number of CI participants. Recruitment was restricted to participants with no residual acoustic hearing (no electro-acoustic stimulation) in the implanted ear.

Fitting these criteria, 18 CI users (5 males) aged 33–76 yr ($\mu = 60.8$ yr, $\sigma = 12.4$ yr) volunteered to take part in this study. Six of these participants already had their F0 and VTL JNDs measured in a previous study (Gaudrain and Başkent, 2018), hence they were asked only to perform the SoS tasks for experiments 1 and 2. Not all 18 participants were able to complete all 3 experiments because of their difficulty: participant P14 was only able to complete experiment 3 (voice JNDs), while participant P17 was only able to complete experiments 2 (SoS comprehension) and 3 (voice JNDs). Thus, in total, out of the 18 CI participants, 16 (aged 41–76 yr, $\mu = 62.1$ yr, $\sigma = 10.9$ yr) took part in experiment 1, 17 (aged 41–76 yr, $\mu = 62.5$ yr, $\sigma = 10.7$ yr) took part in experiment 2, and all 18 took part in experiment 3.

This study was approved by the Medical Ethical Committee of the UMCG (METc 2012.392). All participants were given ample time and information before participation and signed a written informed consent before data collection. All participants were paid an hourly wage for their participation and compensated for their travel costs, as per departmental guidelines.

B. Voice cue manipulations

*F*0 and VTL were manipulated relative to the original voice in each corpus (one corpus per experiment) using STRAIGHT (Kawahara and Irino, 2005). In SoS perception, to prevent the voice manipulation from affecting intelligibility *per se*, the resynthesized voice was always designated as the masker.

In STRAIGHT, *F*0 differences are expressed as a shift in the overall pitch contour by a number of semitones with respect to the average *F*0 of the stimulus. This method helps preserve the fluctuations in the pitch contour of the signal, thus making the synthesized speaker sound more natural (e.g., as was done by Stickney *et al.*, 2007). VTL differences are expressed in STRAIGHT as a compression/stretching in the spectral envelope (formant peaks) of the signal along a linear frequency axis. Shortening VTL results in stretching the spectral envelope toward higher frequencies while

TABLE I. Demographic information for CI users. All durations, in years, are calculated based on the date of testing. Y: yes; N: no; L: left ear; R: right ear. The column "Bimodal user" indicates whether the participant was a bimodal user, and on which ear the hearing aid was. See text for details about the NVA scores. The dynamic range is only provided for cochlear users as the *T*-levels are not routinely measured during fitting sessions of Advanced Bionics (AB, Stäfa, Switzerland) devices. The dynamic range was computed as the mean across all channels of the difference between *C*-levels and *T*-levels in current level units.

Participant	Age (yr)	Processor	Implant	Duration of CI use (yr)	Ear tested	Bilateral user	Bimodal user	Strategy	Duration of hearing loss (yr)	Etiology	Post-operative NVA scores (%)	Dynamic range (current level units)
P04	65.1	Cochlear CP910	CI422	2.6	L	N	N	MP3000	61.6	Meningitis	40	41
P05	65.3	Cochlear CP910	CI24RE CA	6.6	L	N	N	MP3000	13.7	Chronic otitis media	79	79.8
P06	71.0	Cochlear CP910	CI24RE CA	7.7	L	N	N	ACE	60.3	Unknown	90	33.0
P07	52.3	Cochlear CP910	CI24RE CA	8.6	R	N	N	ACE	43.7	Ototoxic medication	48	49.1
P08	76.1	AB Naída Q70	HiRes90k Helix	9.4	R	Y	N	HiRes Optima-S	16.7	Genetic	81	—
P10	52.1	Cochlear CP810	CI24RE CS	14.2	R	N	N	MP3000	31.9	Menière's disease	58	38.8
P12	69.0	Cochlear CP910	CI24R CS	14.5	R	N	N	ACE	23.5	Unknown	90	50.8
P13	75.4	Cochlear CP810	CI24R CA	12.5	R	N	N	ACE	34.9	Unknown	55	58.6
P14	33.3	Cochlear CP810	CI24RE CA	4.0	L	N	N	ACE	29.3	Unknown	48	—
P15	67.9	MedEl Opus 2	MedEl Sonata Medium	3.5	R	N	N	FS4	17.5	Genetic	68	—
P16	68.6	AB Naída Q70	HiRes90k Helix	7.5	R	N	N	HiRes Optima-S	61.1	Unknown	50	—
P17	67.7	Cochlear CP810	Nucleus 24 (CI24M)	16.3	L	N	N	SPEAK	5.4	Chronic otitis media	50	43.5
P18	63.3	AB Naída Q90	HiRes90k HiFocus 1 J	5.8	R	Y	N	HiRes Optima-S	0.2	Genetic	80	—
P19	66.1	AB Naída Q90	HiRes90k HiFocus midscala	0.6	R	N	Y: L	Unknown	19.5	Progressive hearing loss	77	—
P20	67.8	Cochlear CP810	CI24RE CA	3.7	L	N	N	MP3000	47.1	Skull fracture	80	69.9
P21	50.1	AB Neptune	HiRes90k HiFocus 1 J	3.7	R	N	N	HiRes single F120	34.4	Genetic	80	—
P22	41.2	Cochlear CP910	CI422	0.7	R	N	Y: L	MP3000	14.5	Genetic	80	84.1
P23	42.8	AB Naída Q70	HiRes90k Advantage CI HiFocus-1500-04 MS	0.7	R	N	Y: L	HiRes Optima-S	9.1	Osteogenesis imperfecta	95	—

elongating VTL results in spectral envelope compression toward lower frequencies.

Figure 1 shows the [$\Delta F0$,ΔVTL] plane for voice differences relative to the voice of the reference female speaker in experiment 1, shown at the origin of the plane. The dashed ellipses indicate the ranges of relative $F0$ and VTL differences between the reference female voice and 99% of the population based on data from Peterson and Barney (1952). The data from Peterson and Barney were normalized to the average $F0$ (about 176 Hz) and estimated VTL (about 14.4 cm) of the reference female speaker. The reference VTL was estimated following the method of Ives *et al.* (2005), assuming a height of about 170 cm for an average adult Dutch female based on growth curves for the Dutch population (Schönbeck, 2010). ΔVTL is oriented upside down to reflect the fact that negative ΔVTLs translate to an increase in the frequency of the components of the spectral envelope. The red crosses indicate all combinations of $F0$ and VTL manipulations applied in this study relative to the reference female voice. A broad span of $F0$ and VTL differences was chosen to encompass the mean $F0$ and VTL sensitivity thresholds of 9.19 st and 7.19 st, respectively, reported in the literature for CI users (Gaudrain and Başkent, 2018).

Stimuli for all three experiments were sampled at 44.1 kHz, processed, and presented using a custom-built program in MATLAB R2014b (The MathWorks, Natick, MA).

C. Procedure

All experiments were completed in two sessions of 2 h each (including breaks) for CI participants, and in a single session of 2.5 h or less (including breaks) for NH participants. For the CI group, experiment 3 was usually carried out in the first session, while experiments 1 and 2 were completed in the second session. For all experiments, a short training block was provided with feedback to familiarize the participants with the testing procedures.

Bimodal CI users were asked to take off their hearing aids (HAs) during the experiments, and the ear with the HA was plugged. Bilateral users were asked to keep the CI on their better ear and remove the contralateral one. Audiometric measurements without the HA (with ear plugged) and with all CIs removed revealed no residual acoustic hearing (all thresholds were greater than 90 dB HL) for frequencies up to 8 kHz.

All participants were given both oral and written instructions that appeared on an interactive touch screen placed in front of the participant. Participants responded either by tapping a response button on the touch screen (experiments 2 and 3) or verbally (experiment 1).

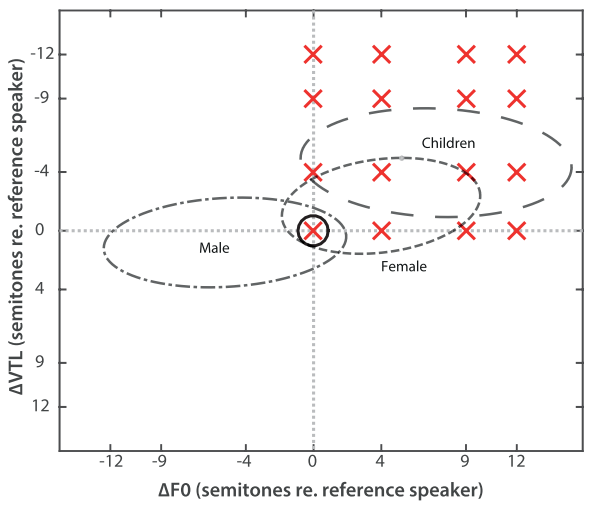

FIG. 1. (Color online) [$\Delta F0$, ΔVTL] plane. The reference female speaker is at the origin of the plane, as indicated by the solid circle. Decreasing $F0$ and elongating VTL yields deeper-sounding male-like voices, while increasing $F0$ and shortening VTL yields child-like voices. Dashed ellipses, derived from the data of Peterson and Barney (1952), indicate the ranges of typical $F0$ and VTL differences between the reference female speaker from experiment 1 and 99% of the population. The data of Peterson and Barney were normalized to the reference female speaker in experiment 1. The red crosses indicate the 16 different combinations (experimental conditions) of $\Delta F0$ and ΔVTL used in both experiments 1 and 2.

D. Apparatus

All experiments were conducted in a soundproof anechoic chamber. The processed stimuli were presented via an AudioFire4 soundcard (Echo Digital Audio Corp, Santa Barbara, CA) connected through Sony/Philips Digital Interface (S/PDIF) to a DA10 D/A converter (Lavry Engineering, Poulsbo, WA) and a Tannoy loudspeaker (Tannoy Precision 8D; Tannoy Ltd., North Lanarkshire, UK), placed 1 m away from the participant.

III. EXPERIMENT 1: THE EFFECT OF $\Delta F0$ AND ΔVTL ON SOS INTELLIGIBILITY

A. Rationale

This experiment, along with experiment 2, was designed to answer the first research question posed in this study, which is whether CI users, similar to NH listeners, could benefit from increasing $\Delta F0$ and ΔVTL between target and masker voices in a SoS sentence intelligibility task. SoS intelligibility scores were measured as a function of systematically increasing $\Delta F0$ and ΔVTL between the target and masker speakers.

B. Methods

1. Stimuli

Stimuli were taken from the corpus of Dutch sentences (e.g., "*Buiten is het donker en koud*" [*Outside it is dark and cold*]) created by Versfeld *et al.* (2000). Versfeld *et al.* collected sentences from large databases, such as Dutch newspapers, following the procedures highlighted by Plomp and Mimpen (1979). From this initial collection of sentences, Versfeld *et al.* selected those that had neutral semantic content and were syntactically and grammatically correct. The final selection of sentences was divided into 39 lists of 13 phonemically balanced sentences. In this experiment, all sentences were chosen from the female speaker in the corpus who had an average $F0$ of 176 Hz.

Target sentences were taken from lists 1–12 and 15–18 (for a total of 16 lists; 1 list per condition), and training sentences were taken from list 14. List 13 contained repetitions

from list 21 (Clarke *et al.*, 2014), while list 39 did not match the average frequency distribution of phonemes in Dutch (Versfeld *et al.*, 2000). Hence, these three lists were used for constructing the masker.

All sentences in the corpus designated for use as maskers were first processed offline using STRAIGHT with all 16 combinations (experimental conditions) of $F0$ and VTL differences, as shown in Fig. 1. For the condition $\Delta F0 = 0$ and $\Delta\text{VTL} = 0$, the masker was still processed with STRAIGHT, with no change in $F0$ or VTL introduced. The target speaker was always kept as the original female in the corpus and not processed with STRAIGHT, and all target sentences were equalized in intensity to the same root-mean-square (RMS) value.

In each trial, the masking sentence sequence was designed to start 500 ms before the onset of the target sentence and end 250 ms after the offset of the target. The masking sentence sequence was built by randomly choosing 1-s-long segments from the STRAIGHT-processed masker sentences with the given $\Delta F0$ and ΔVTL combination associated with the given trial. A raised cosine ramp of 2 ms was applied both to the beginning and end of each segment. All segments were then concatenated, and the masker was trimmed to an appropriate duration. This procedure yielded maskers that were partly intelligible but were not grammatically or semantically meaningful as a sentence. Finally, 50-ms raised cosine ramps were applied both to the beginning and end of the entire masker sequence.

The target speech was calibrated to 65 dB sound pressure level (SPL). The RMS of the entire masker sequence was adjusted to achieve the target-to-masker ratio (TMR) of +8 dB for CI and −8 dB for NH groups. The TMR values for both groups were chosen to obtain a performance between 40% and 60% based on pilot data collected for this experiment at various TMRs. To help the participants familiarize themselves quickly with the task, the TMR used for the training block was 4 dB higher than the one used during actual testing (i.e., set at +12 dB for the CI group and −4 dB for the NH group).

2. Procedure

This task aimed to measure speech intelligibility of the target sentence. Participants were always presented with a single target-masker combination in a given trial and asked to focus on the target sentence, which started 500 ms after the masker. They were asked to repeat anything they heard, even if they thought it made no sense or if what they heard was only a single word or part of a word.

Participants were given a short training block consisting of 12 sentences randomly selected from the 13 available in the training list. Six of these sentences were presented first in quiet to familiarize the participants with the target female speaker, and then the remaining six were presented with a competing masker to familiarize participants with the actual experimental procedure. The [$\Delta F0$,ΔVTL] values for this competing talker were both set to [+8 st,−8 st]. This combination was not present during actual testing so as not to bias the experimental results but was sufficiently large for most CI participants to be able to detect the voice difference between the target and masker. During training (in quiet and in noise), both auditory and visual feedback were given after the participant's response, such that the correct target sentence was shown on the screen while the entire stimulus was played a second time through the loudspeaker.

The actual test was comprised of a total of 208 trials (13 sentences per list × 16 conditions). All 208 stimuli were generated offline before the experiment began and presented in a pseudo-randomized order to each participant. No feedback was given during actual testing: participants only heard the stimulus once, gave their verbal response, and were not shown the correct target sentence on the screen.

The verbal responses were scored online on a word-by-word basis using a graphical user interface (GUI) implemented in MATLAB. For each correctly repeated word, the experimenter would click its corresponding button on the scoring GUI, which was not seen by the participant. A similar GUI was also developed and used for offline scoring of the responses. Online scoring was performed during data collection by a native Dutch-speaking student assistant to minimize potential misinterpretation of the CI users' articulation. In addition, the vocal responses from the participants were recorded and offline scoring was performed after data collection to double-check that no word was incorrectly scored during the online scoring.

A response word was considered correct even if some minor confusions were made, such as confusing different forms of the same personal pronoun (e.g., saying "*zij*" instead of "*ze*" [*she*] or "*wij*" instead of "*we*" [*we*]), confusing the words "*this*" and "*that*," "*shall*" and "*should*," "*can*" and "*could*," using the diminutive form (e.g., saying "*hondje*" instead of "*hond*" [*puppy* vs *dog*]), or repeating the words in a different order than the one in the target sentence. Repeating additional words that were not in the target was not penalized.

A response word was considered incorrect if part of the word was repeated instead of the full word (e.g., saying "*kast*" instead of "*koelkast*" [*cupboard* vs *fridge*]), an extra addition was made to the word (e.g., saying "*zeiltocht*" when the actual word was "*tocht*" [*sailing trip* vs *trip*]), tenses were confused (e.g., past and present), singular and plural were confused, or pronouns were confused (e.g., saying "*she*" instead of "*he*"). Responses were not checked as to whether they matched some of the masking words.

A total of four scheduled breaks were programmed into the experiment script, however, participants were told to request additional breaks whenever they needed, and the experimenter could also decide on a break if she felt that a participant was becoming tired. The entire experiment (training, test, and breaks) was completed within 1.5 h.

3. Apparatus

Participants' verbal responses were recorded for offline analyses using a RØDE NT1-A microphone mounted on a RØDE SM6 with pop-shield (RØDE Microphones LLC, Silverwater, Australia). The microphone was connected to a PreSonus TubePre v2 amplifier (PreSonus Audio Electronics, Inc., Baton Rouge, LA), which was connected to the Apple Mac computer (Apple Inc., Cupertino, CA)

running MATLAB R2014b via an AudioFire soundcard (Echo Digital Audio Corp, Santa Barbara, CA). The recording started automatically with the onset of the stimulus via the experiment script in MATLAB. All recordings were stored as FLAC (free lossless audio codec) files with a sampling rate of 44.1 kHz.

4. Statistical analyses

All data in this study were analyzed using *R* (version 3.3.3, R Core Team, 2017), and linear modeling was done using the *lme4* package (version 1.1-15, Bates *et al.*, 2015).

To quantify the effect of each of the *F*0 and VTL differences on the SoS intelligibility score, a generalized linear mixed-effects model (GLMM), with a logit link function, was fitted to the binary per-word score using the following equation in *lme4* syntax:

$$\text{score} \sim f0 * vtl * \text{group} + (1 + f0 * vtl|\text{participant}). \quad (1)$$

The fixed effect term $f0 * vtl *$ group indicates the full factorial model, including each main effect and all interactions. The terms *f*0 and *vtl* are the normalized versions of $\Delta F0$ and ΔVTL, respectively, and are defined by $f0 = \Delta F0/12$ and $vtl = \Delta\text{VTL}/12$. The term *group* refers to the participant group: NH or CI. The term $(1+f0 * vtl|\text{participant})$ defines a random intercept and slope per participant for each of *f*0, *vtl*, and the interaction term, making the model comparable to a repeated-measures analysis of variance (ANOVA). The GLMM described by Eq. (1) was used to look at the overall effect of group, and whether $\Delta F0$ and ΔVTL had significantly different effects per group. The coefficients for each factor of the model, its associated Wald's *z*-value, and its corresponding *p*-value are reported.

The following GLMM was fitted to determine the effect of $\Delta F0$ and ΔVTL on SoS intelligibility scores for each group separately

$$\text{score} \sim f0 * vtl + (1 + f0 * vtl\,|\,\text{participant}). \quad (2)$$

This is the same as the model in Eq. 1, but without the group effect. The random slopes represent the respective weights of *f*0 and *vtl* per participant for this task, expressed in the logistic regression function as:

$$\text{logit}(\text{score}) = a \cdot f0 + b \cdot vtl + c \cdot (f0 \times vtl) + d. \quad (3)$$

In Eq. (3), *a* is the participant-specific slope (weight) for *f*0, *b* is the participant-specific slope for *vtl*, *c* is the participant-specific slope for the interaction term $f0 \times vtl$, and *d* is the intercept per participant.

C. Results

Figure 2 shows the average SoS intelligibility scores per group for each condition of $\Delta F0$ and ΔVTL. The SoS intelligibility score, in percent, is defined as the number of correctly repeated words divided by the total number of words in all target sentences presented per condition.

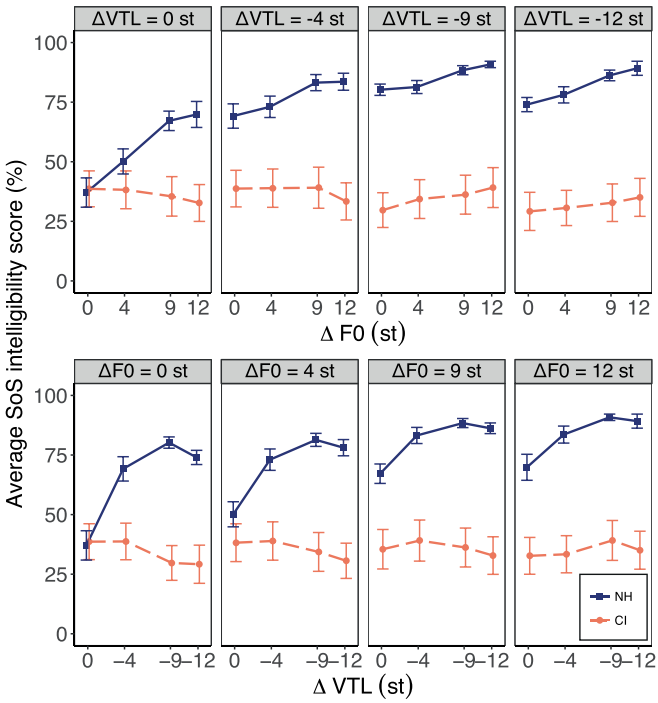

FIG. 2. (Color online) SoS percent-correct intelligibility scores averaged per group for each condition of $\Delta F0$ and ΔVTL. Dark squares with solid lines represent the NH data, while light circles with dashed lines represent the CI data. Error bars represent one standard error from the mean. (Top row) SoS intelligibility scores plotted as a function of increasing $\Delta F0$ between target and masker speakers for each value of ΔVTL, as indicated by the individual panels. (Bottom row) SoS intelligibility scores plotted as a function of increasing ΔVTL for each value of $\Delta F0$, as indicated by the individual panels.

The data show that for the NH group, SoS intelligibility increased as a function of increasing the voice cue difference ($\Delta F0$, ΔVTL, or both) between the target and masker speakers. In contrast, the CI group showed no benefit in SoS intelligibility from increasing $\Delta F0$, in addition to a slight decrement in SoS intelligibility as a function of increasing ΔVTL.

1. Between-group effects

Between-group effects were analyzed first to confirm whether the starting SoS intelligibility level for both participant groups under the baseline condition [$\Delta F0 = 0$, $\Delta\text{VTL} = 0$] was comparable. The coefficients obtained from the logistic regression (provided in Table II) revealed no effect of group for this baseline condition where the target and masker voices belonged to the same speaker. This confirms that the TMR chosen for each group from pilot data did succeed in equating the baseline performance of the two groups.

The logistic regression model also revealed a significant effect of both $\Delta F0$ and ΔVTL on SoS intelligibility. However, the type of this effect (benefit or decrement in intelligibility) was different for each group, as indicated by the significant interaction between $\Delta F0$ (ΔVTL) and participant group. Finally, different combinations of $\Delta F0$ and ΔVTL did not lead to the same degree of benefit in SoS intelligibility across groups, as indicated by the significant interaction between group, $\Delta F0$, and ΔVTL.

TABLE II. Coefficients obtained from the logistic regression [Eq. (3)] with the normalized variables *f*0 and *vtl*. β represents the estimated value of the coefficient, SE represents the standard error of that estimate, z is the Wald z-statistic, and p represents its corresponding p-value. Significance codes: $p < 0.05$ '*'; $p < 0.01$ '**'; $p < 0.001$ '***'.

Fixed effect coefficient	Overall effect of group	NH group	CI group
Intercept	$\beta = -0.20$, SE = 0.34, $z = -0.58$, $p = 0.56$	$\beta = -0.20$, SE = 0.23, $z = -0.86$, $p = 0.39$	$\beta = -0.63$, SE = 0.46, $z = -1.35$, $p = 0.18$
*f*0	$\beta = 1.44$, SE = 0.17, $z = 8.58$, $p < 0.001$***	$\beta = 1.44$, SE = 0.16, $z = 8.86$, $p < 0.001$***	$\beta = -0.61$, SE = 0.20, $z = -3.00$, $p = 0.003$**
vtl	$\beta = 1.76$, SE = 0.17, $z = 10.24$, $p < 0.001$***	$\beta = 1.75$, SE = 0.15, $z = 11.56$, $p < 0.001$***	$\beta = -1.02$, SE = 0.23, $z = -4.50$, $p < 0.001$***
group	$\beta = -0.44$, SE = 0.50, $z = -0.87$, $p = 0.38$	—	—
*f*0 × *vtl*	$\beta = -0.48$, SE = 0.22, $z = -2.13$, $p = 0.03$*	$\beta = -0.48$, SE = 0.19, $z = -2.51$, $p = 0.012$*	$\beta = 1.22$, SE = 0.31, $z = 3.98$, $p < 0.001$***
*f*0 × group	$\beta = -2.05$, SE = 0.26, $z = -7.93$, $p < 0.001$***	—	—
vtl × group	$\beta = -2.73$, SE = 0.27, $z = -10.26$, $p < 0.001$***	—	—
*f*0 × *vtl* × group	$\beta = 1.68$, SE = 0.35, $z = 4.82$, $p < 0.001$***	—	—

2. NH listeners

The effects of $\Delta F0$ and ΔVTL on SoS intelligibility were analyzed separately for each group. For the NH listeners, the logistic regression model revealed that SoS intelligibility improved by 0.17 Berkson[2] (Bk)/st increase in $\Delta F0$ and by 0.21 Bk/st increase in ΔVTL. The size of the benefit in SoS intelligibility from increasing $\Delta F0$ was found to depend on the value of ΔVTL, as indicated by the significant interaction between $\Delta F0$ and ΔVTL. This effect can be seen in the top panel of Fig. 2, such that for certain values of ΔVTL, NH participants were likely to gain larger improvements in SoS intelligibility from increasing $\Delta F0$.

The participant-specific slopes (weights), which are the subject-specific mixed-effects deviation from the fixed group estimate for the normalized coefficients *f*0 and *vtl*, are provided in Table III. Notice that the slopes for *f*0 and *vtl* are positive for all NH participants, indicating that SoS

TABLE III. Subject-specific weights (subject-specific mixed-effects deviation from the fixed group estimate) for the normalized terms *f*0, *vtl*, and the interaction effect. Here, *f*0, *vtl*, and the interaction term refer to the coefficients *a*, *b*, and *c*, respectively, in the logistic regression function, while the intercept refers to *d*.

NH					CI				
Participant	Intercept	*f*0	*vtl*	*f*0 × *vtl*	Participant	Intercept	*f*0	*vtl*	*f*0 × *vtl*
NH-P02	− 0.31	0.34	1.98	0.27	CI-P04	− 3.79	− 0.82	− 1.64	1.52
NH-P03	1.51	0.92	1.76	− 0.75	CI-P05	0.02	0.40	− 0.56	0.34
NH-P04	− 0.50	1.79	2.25	− 1.11	CI-P06	− 0.14	− 0.46	− 0.66	1.70
NH-P05	− 1.57	0.92	2.55	− 1.00	CI-P07	− 3.09	− 0.96	− 1.34	2.30
NH-P06	1.04	1.32	1.12	− 0.42	CI-P08	0.81	− 0.29	− 0.88	1.20
NH-P07	− 1.87	2.14	2.23	− 0.80	CI-P10	− 3.66	− 1.05	− 1.38	1.60
NH-P08	0.30	1.95	1.34	− 0.63	CI-P12	1.04	− 0.19	− 0.46	0.20
NH-P09	− 0.35	1.37	1.62	− 0.38	CI-P13	0.49	− 0.20	− 0.93	0.36
NH-P10	− 1.03	1.76	1.77	− 0.70	CI-P15	− 1.82	− 1.43	− 2.07	1.78
NH-P11	− 1.19	1.01	2.37	0.49	CI-P16	− 1.62	− 0.07	0.10	0.20
NH-P12	0.97	1.02	0.78	0.31	CI-P18	0.16	− 1.03	− 1.42	2.04
NH-P13	0.12	1.08	1.55	− 0.24	CI-P19	− 0.83	− 1.24	− 1.98	2.26
NH-P14	− 0.93	2.45	1.79	− 0.43	CI-P20	− 1.13	− 0.51	− 0.89	0.25
NH-P15	0.11	1.41	2.16	− 0.86	CI-P21	− 0.79	− 1.41	− 1.68	2.44
NH-P16	− 1.19	2.30	2.12	− 1.06	CI-P22	2.64	− 1.08	− 0.95	1.11
NH-P17	0.17	1.43	1.87	− 0.57	CI-P23	1.85	0.61	0.54	0.11
NH-P18	0.75	1.84	0.80	− 0.76					
NH-P19	0.43	0.72	1.28	0.03					
Minimum	− 1.87	0.34	0.78	− 1.11	Minimum	− 3.79	− 1.43	− 2.07	0.11
Maximum	1.51	2.45	2.55	0.49	Maximum	2.64	0.61	0.54	2.44
Mean	− 0.20	1.43	1.74	− 0.48	Mean	− 0.62	− 0.61	− 1.01	1.21
Standard deviation	0.96	0.58	0.52	0.48	Standard deviation	1.87	0.62	0.71	0.85

intelligibility improved as a function of increasing $\Delta F0$ and ΔVTL between target and masker.

3. CI listeners

In contrast to the NH group, who showed a benefit from increasing both $\Delta F0$ and ΔVTL between target and masker voices, the CI group revealed a significant decrement in SoS intelligibility of about 0.07 Bk/st increase in $\Delta F0$ and a decrement of about 0.12 Bk/st increase in ΔVTL. This finding contradicts the hypothesis that increasing $\Delta F0$ and ΔVTL between target and masker voices should lead to an improvement in SoS intelligibility for CI users. The significant interaction term reveals that the detrimental effect of increasing $\Delta F0$ and ΔVTL on SoS intelligibility changes according to the combination of $\Delta F0$ and ΔVTL. As shown in the top panels of Fig. 2, increasing $\Delta F0$ between target and masker was detrimental for SoS intelligibility until ΔVTL was -4 st. When ΔVTL was -9 st and -12 st, increasing $\Delta F0$ led to a slight improvement in SoS intelligibility, although this improvement did not turn out to be significant when the logistic regression was applied only for ΔVTL values larger than -4 st [$\beta = 1.22$, standard error (SE) $= 0.88$, $z = 1.39$, $p = 0.17$].

D. Discussion

The first research question in this study was whether CI users would benefit from $F0$ and VTL differences between target and masker speakers in a SoS intelligibility task similar to NH listeners. To explore this question, in this experiment, $F0$ and VTL of the masker speaker were manipulated relative to the voice of the original female speaker (target). The effect of increased voice differences on SoS was explored by measuring intelligibility as a function of increasing $\Delta F0$ and ΔVTL between target and masker for both NH and CI users.

NH listeners gained an improvement (benefit) in SoS intelligibility scores as a function of increasing $\Delta F0$ and/or ΔVTL of the masker relative to those of the target speaker, which is consistent with the effects reported in a number of studies (e.g., Assmann and Summerfield, 1990; Başkent and Gaudrain, 2016; Darwin *et al.*, 2003; Vestergaard *et al.*, 2009). In contrast, CI users demonstrated a slight but significant decrement in SoS intelligibility with increasing $\Delta F0$ and/or ΔVTL between target and masker speakers. Because the target in the current experiment always remained the same voice in all conditions, this decrease in intelligibility with an increase in $\Delta F0$ or ΔVTL is akin to increasing the influence of the masker. The literature reports mixed findings for CI users regarding the benefit from $F0$ differences between target and masker speakers, either manipulated from the same talker, as was done here, or by use of different speakers with differing $F0$s. While Stickney *et al.* observed no improvement in SoS scores for CI users, either when the masker sentence was from a different talker (Stickney *et al.*, 2004) or when the masker voice was the same talker as the target with its $F0$ manipulated (Stickney *et al.*, 2007), Pyschny *et al.* (2011) reported a systematic benefit in a similar condition.

One fundamental difference between the studies of Stickney *et al.* and Pyschny *et al.* (2011) is that the CI users recruited in the latter study were all bimodal users. These bimodal CI users, even though tested without their HAs, had presumably sufficient residual acoustic hearing that may have helped them draw a benefit from $F0$ differences in SoS. In fact, previous literature has reported that low-frequency acoustic cues in residual hearing, even when limited, can help preserve $F0$ cues to a large extent, enhancing the sensitivity to such cues (Başkent *et al.*, 2018). In addition, perhaps as a result of their residual acoustic hearing, these CI users were able to perform the SoS task at a TMR that was unusually low for CI users (0 dB), and still managed to produce SoS scores that were well above floor performance, varying between roughly 30% and 45%. It has been shown that the amount of benefit from voice cue differences between target and masker speakers highly depends on the TMR tested (e.g., Darwin *et al.*, 2003; see Figs. 4 and 8 in Stickney *et al.*, 2004): at high TMRs, the benefit from increasing $F0$ or VTL between target and masker speakers becomes minimal, which may be related to placing more emphasis on loudness cues from the target compared to voice cue differences between the two talkers in a SoS task. In comparison to the bimodal CI participants tested by Pyschny *et al.* (2011), the CI users tested by Stickney *et al.* (2004) and Stickney *et al.* (2007) could not reach the same level of high performance, even when tested at a relatively high TMR (above $+10$ dB). Because the CI participants tested in the present study were recruited to have a wide range of speech-in-quiet intelligibility scores, they were all tested at a relatively high TMR of $+8$ dB, similar to both Stickney *et al.* studies. Thus, the positive effect of increasing $\Delta F0$ on SoS intelligibility observed by Pyschny *et al.* may be limited to high-performing bimodal participants who may have access to residual acoustic cues, including $F0$ cues, even without their HAs. This may allow them to be tested at low TMRs, where the interactive effects may be stronger than at high TMRs. With that said, because the TMR has been shown to play an important role in the amount of benefit from voice differences between two competing talkers, the difference in the patterns of performance between NH and CI listeners could be attributed to the different TMRs used to test each group. Thus, the systematic effect of TMR on the benefit from voice cue differences in SoS tasks for both NH and CI users should be investigated in a future study.

Data from this experiment revealed that, contrary to what was expected, increasing the masker's $F0$ and shortening its VTL relative to the target voice (toward a child-like voice) appeared to increase the masking effect for the CI group. This effect has been previously reported in the literature by Pyschny *et al.* (2011), where they observed a decrement in CI user's performance as they increased ΔVTL. As was done in the current study, Pyschny *et al.* also manipulated the masker along the direction of shorter VTLs relative to the target. The authors attributed this adverse effect of ΔVTL to the masker being more salient than the target because of its shorter VTL. A similar effect was also reported for both NH and CI listeners in a study by Cullington and Zeng (2008), in which they observed a stronger masking effect of child maskers compared to female

maskers when the target was a male speaker. This is counter-intuitive because, in principle, the $F0$ and VTL differences between a child and an adult male speaker are usually larger than those between an adult female and an adult male (Peterson and Barney, 1952; Smith and Patterson, 2005).

A possible explanation for this effect in CI users is provided by Fig. 3, which shows the effect of increasing $\Delta F0$ and ΔVTL between target and masker speakers on the resulting TMR per simulated CI electrode and electrodogram patterns. Figure 3(A) shows the TMR per electrode averaged across all target sentences used in this experiment, with masker combinations obtained as described in Sec. III B 1. The top part of Fig. 3(A) shows the TMR computed for only increasing $F0$ of the masker relative to that of the target. As $F0$ increases, the TMR appears to decrease, especially along the higher frequencies (electrodes 1–14). The bottom part of Fig. 3(A) shows the effect of shortening the masker's VTL relative to that of the target. As the masker's VTL is shortened, the TMR decreases dramatically for the lower frequency components of the stimuli (electrodes 12–22), indicating an effective increase in masking effect. Figure 3(B) demonstrates this effect on the stimulation pattern using a sample stimulus. For $F0$ differences [top part of Fig. 3(B)], the masker (bright) and target (dark) patterns do not appear to change dramatically. However, for VTL differences between masker and target, the masker pattern appears to stretch along higher frequencies, spreading to higher-frequency channels (represented by electrodes 16–22). This happens because shortening VTL leads to a stretching of the spectral envelope along a linear frequency scale toward higher frequencies, as can be seen in Fig. 3(C), which shows the spectrograms of the maskers before being processed by the CI simulation. Hence, when shortening the masker's VTL by 12 st, the lower frequencies of the target become completely masked, compared to the case when ΔVTL was 0 st. This is because these low-frequency patterns of the masker start occupying more of the same low-frequency channels as those of the target, leading to the fusion of masker and target components in that frequency range. Thus, as ΔVTL increases in this experiment, a stronger masking effect can be expected for the CI group.

In the following experiment, a different task was administered to measure the effect of voice cue differences between competing speakers on another aspect of SoS perception, namely, SoS comprehension. Sentence comprehension was assessed in the following experiment because it more closely mimics real-life communication scenarios (Best *et al.*, 2016) in which listeners extract meaningful information from the incoming sentence and formulate the appropriate response accordingly (Rana *et al.*, 2017). In addition, it is a process that taps into higher levels of cognitive processing. According to Kiessling *et al.* (2003), "Comprehending is an activity undertaken beyond the processes of hearing and listening [and] is the perception of information, meaning or intent." Thus, when the acoustic signal is impoverished, as is the case with CI processing,

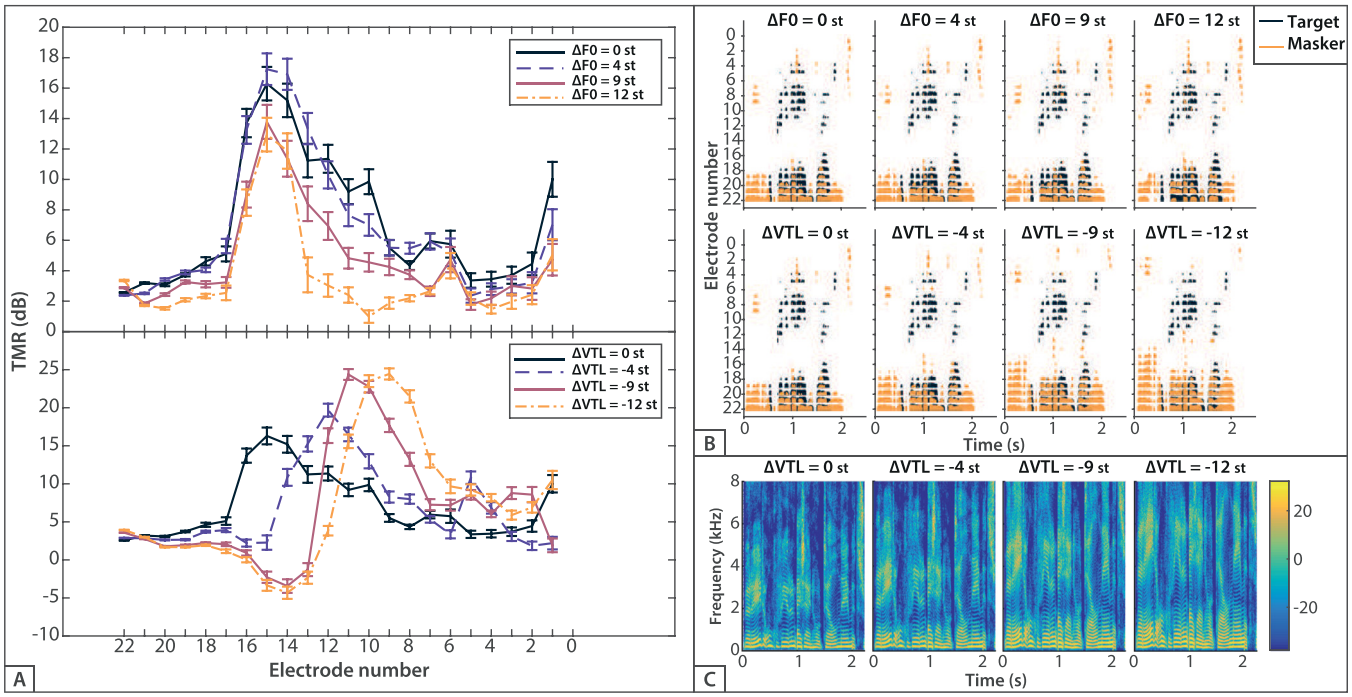

FIG. 3. (Color online) Effect of increasing $\Delta F0$ and ΔVTL between target and masker speakers in simulations of CI processing using the Nucleus MATLAB Toolbox (NMT version 4.31; Swanson and Mauch, 2006) from Cochlear. (A) (top) TMR per electrode averaged across the entire speech corpus for only changing $\Delta F0$. Error bars indicate one standard deviation from the mean TMR. (A) (bottom) Same as the top panel but for changes in only ΔVTL. (B) (top row) Electrodograms obtained for a sample stimulus using NMT, with fixed target sentence "*We kunnen weer even vooruit*" [*We can move forward again*], and identical masker mixture at a TMR of +8 dB. Only $\Delta F0$ is varied and ΔVTL is kept at 0 st. Dark patterns indicate the pattern produced by the target, while bright patterns indicate that of the masker. (B) (bottom row) Same as the top panel, but for changes in only ΔVTL, while $\Delta F0$ is kept at 0 st. (C) Spectrograms obtained for the same maskers as in the bottom row of (B) (only ΔVTL varied while $\Delta F0$ kept at 0 st) before processing with NMT.

overall sentence comprehension may be compromised if CI users cannot understand a sufficient number of words to draw meaning from the entire sentence. This would not be evident in a typical sentence intelligibility task, since the CI users may repeat a number of words per sentence, but these words could be insufficient in helping them assign meaning to the sentence. In addition, sentence comprehension speed (RTs) could also be easily assessed, which has been shown in the literature to capture more robust effects of task difficulty compared to traditional accuracy measures (e.g., Baer *et al.*, 1993; Gatehouse and Gordon, 1990; Hecker *et al.*, 1966). Such RTs could not have been easily measured using a task as that deployed in experiment 1.

IV. EXPERIMENT 2: EFFECT OF $\Delta F0$ AND ΔVTL ON SOS COMPREHENSION USING A SVT

A. Rationale

SoS comprehension as a function of $\Delta F0$ and ΔVTL between two competing talkers was assessed in this experiment using a Dutch SVT (see Adank and Janse, 2009, for a description). Based mainly on the English speed and capacity of language processing task (Baddeley *et al.*, 1992), the Dutch SVT is comprised of true and false sentence pairs, which allows for measuring not only verification (comprehension) accuracy but also RTs. Because differences across experimental conditions were shown to manifest more robustly using RTs than using traditional accuracy (percent-correct) scores alone (e.g., Baer *et al.*, 1993; Gatehouse and Gordon, 1990; Hecker *et al.*, 1966), RTs have been extensively used in the literature as an additional measure of performance. For example, adverse listening conditions require a relatively longer time to process and thus lead to longer RTs, compared to ideal listening conditions (Baer *et al.*, 1993; Gatehouse and Gordon, 1990).

While SVT provides two measures, one accuracy and the other speed of comprehension, it is often challenging to interpret accuracy and RT measures in isolation, since a participant may, for example, respond at a slower rate at the expense of higher accuracy (e.g., Pachella, 1974; Schouten and Bekker, 1967; Wickelgren, 1977). This speed-accuracy trade-off can be addressed by combining accuracy and RT measures into a unified measure of performance called the *drift rate* (for a review, see Ratcliff *et al.*, 2016), which represents the rate of evidence accumulation to reach a decision (labeling the sentence as true or false). This measure can provide insight into the quality of information gathered by the participant across different experimental conditions, and is assumed to be appropriate for measuring task difficulty (Wagenmakers *et al.*, 2007), such that a slower drift rate would indicate a more difficult task.

In this experiment, the drift rate was computed using the EZ-diffusion model provided by Wagenmakers *et al.* (2007), which is a simplified version of the full drift-diffusion model introduced by Ratcliff (1978). The EZ-diffusion model makes use of the RT distribution (both mean and variance) to correct responses, along with the accuracy score to compute the drift rate. Following the method of Wagenmakers *et al.* (2007), the assumptions permitting the use of this model were all satisfied when checked on the data collected.

1. Stimuli

The same masker-target conditions used in experiment 1 were used here with the same 16 combinations of $\Delta F0$ and ΔVTL and at the same TMRs for each group. The sentences used to construct the masker sequences were also the same as in the previous experiment. The only difference between the setup of this experiment and that of the previous one is that, here, the target sentences were taken from Adank and Janse (2009) to obtain both accuracy and RT measures. The Dutch SVT corpus of Adank and Janse contains 100 pairs of sentences, and each pair is comprised of the true (e.g., *Bevers bouwen dammen in de rivier [Beavers build dams in the river]*) and false (e.g., *Bevers grooien in een moestuin [Beavers grow in a vegetable patch]*) versions of a given sentence. The sentences are all grammatically and syntactically correct.

a. Recording of SVT material. Because manipulation of the masker's *F*0 and VTL relative to those of the target was of interest here, it was essential to have the target and masking sentences uttered by the same speaker. Hence, both the sentences from the Dutch SVT and the sentences by Versfeld *et al.* used as maskers (lists 13, 21, and 39) were re-recorded from a native Dutch female speaker, with an average *F*0 of 188 Hz. The Dutch speaker was a 25-yr-old female from the northern provinces of the Netherlands.

The recordings were done in a sound-isolated anechoic chamber using a RØDE NT1-A microphone mounted on a RØDE SM6 with pop-shield (RØDE Microphones LLC, Silverwater, Australia) connected to a PreSonus TubePre v2 preamplifier (PreSonus Audio Electronics, Inc., Baton Rouge, LA). The preamplifier output was connected to the left channel of a DR-100 MKII TASCAM recorder (TEAC Europe GmbH, Wiesbaden, Germany), by which recordings were captured at a sampling rate of 44.1 kHz.

All 200 true/false sentences were recorded three times, with sentences being presented in a randomized order. The best of three recordings was chosen and equalized in RMS. Clicks were smoothed out to decrease noise and pauses longer than 250 ms were shortened to 250 ms.

In addition to the 200 true/false sentences, 8 more true/false sentences were developed and recorded by the same female speaker to be used for training (see Appendix A).

2. Procedure

In this experiment, participants were instructed to indicate whether the target sentence was true or false by pressing the corresponding button on the touchscreen and were requested *not* to repeat the sentence. They were asked to give the first response that came to mind without overthinking.

It is important to note that the Dutch SVT developed by Adank and Janse (2009) is not divided into lists as was done in the English SVT developed by Baddeley *et al.* (1995). The Dutch and aforementioned English SVTs are also slightly different than the SVT developed by Pisoni *et al.* (1987), such that the resolving word, which determines whether the statement is true or false, is not always at the end of the sentence, as is the case in the SVT developed by

Pisoni *et al*. This has potential consequences on measuring RTs as such measurements are usually marked starting from the offset of the resolving word. In the original design of Adank and Janse (2009) negative RTs were possible since the resolving word was not always at the end of the stimulus sentence. Here, however, participants were only able to respond after the offset of the entire stimulus; therefore, negative RTs were not allowed. Nonetheless, the issue of not having the resolving word at the end of the sentence was addressed in the analyses because it could have potentially contributed to the variability in the RTs measured.

The design of this experiment was further modified to accommodate the CI participants. This involved not implementing a timeout window for collecting responses, and not giving speed instructions. These modifications, which were similar to those done by Gatehouse and Gordon (1990) with their hearing impaired participants, were introduced so as not to stress the CI participants who already experience reduced spectrotemporal acoustic-phonetic details of speech, and hence, may end up sacrificing accuracy for speed.

Training was provided in two parts to familiarize participants with the task. In the first part, two true/false sentence pairs from the training list were presented in quiet. In the second part, the remaining two true/false sentence pairs from the training list were presented in the presence of a competing masker at a TMR 4 dB higher than that used during data collection. The voice of the masker differed from that of the target by a $\Delta F0$ of +8 st and a ΔVTL of −8 st.

During actual testing, the first 192 sentences (12 sentences per condition × 16 conditions) from the overall 200 true/false sentences were chosen as the target sentences. For a given condition, 6 true and 6 false sentences were randomly chosen from the 192, with no true/false pair assigned to the same [$\Delta F0$, ΔVTL] condition. All 192 stimuli were generated offline before the experiment began and presented in a pseudo-randomized order to each participant.

Feedback was only provided during training: participants received both auditory and visual feedback for both parts of the training: the target sentence was displayed on the screen, along with whether it was true or false, and the whole stimulus was repeated through the loudspeaker. The entire experiment lasted a maximum of 1 h (including breaks).

3. Statistical analyses

Accuracy scores were converted into the sensitivity measure d' (Green and Swets, 1966) because percent correct responses may be prone to a participant's bias for choosing a specific response for all items. The d' and drift rate data were fit using a linear mixed-effects model (using *lmer* function in *R*), with the same parameters as outlined in Sec. III B 4. $\Delta F0$ and ΔVTL were also normalized as in experiment 1.

Because no timeout was implemented and no speed instructions were given to the participants, RTs above 6 s were discarded (assigned as an incorrect response), and only those RTs corresponding to correct responses were analyzed. The discarded RT measurements amounted to 0.74% of the NH data and 3.16% of the CI data.

Because RT data are positively skewed, they were fit using a GLMM following the recommendations provided by Lo and Andrews (2015), where the effect of the stimulus item (sentence) was included as a random factor [$(1|\text{item})$ term]. This term was introduced to address the potential variability in RTs arising from the issue that the resolving word was not always at the end of the sentence. The resulting model for RTs was of the form $-1/\text{RT}\,(\text{s}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$, where x_i represents the ith fixed effect, and β_i is the corresponding coefficient.

B. Results

Figure 4 shows the mean accuracy scores in d' (top row), the mean RTs (middle row), and the mean drift rate

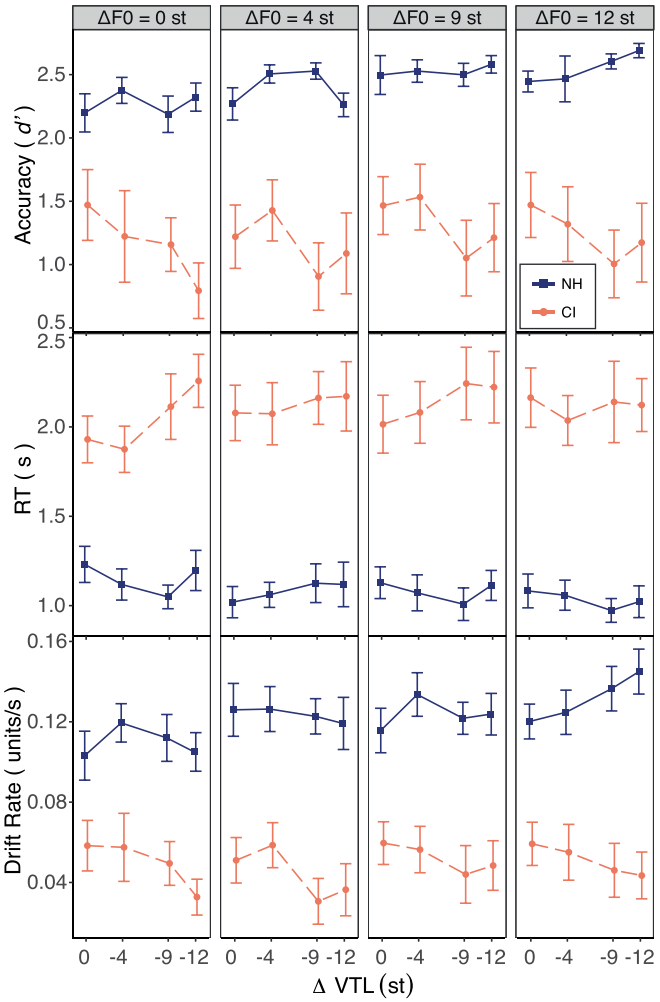

FIG. 4. (Color online) SoS comprehension performance, measured using SVT, averaged per group for each condition of $\Delta F0$ (different panels) and ΔVTL (x axis). Dark squares with solid lines represent the NH data, while light circles with dashed lines represent the CI data. Error bars represent one standard error from the mean. (Top row) SoS comprehension accuracy measured in d'. (Middle row) SoS comprehension RTs measured in seconds. (Bottom row) SoS comprehension drift rate measured in arbitrary units per second.

(bottom row) as the combined measure of performance from both accuracy and RT data.

1. Between-group effects

The regression models for the between-group effects for each of the RTs and drift rate were simplified to exclude the random slopes estimated per participant, since the simplified models did not significantly differ from the full models ($p > 0.13$). The regression model for d' was also simplified in the same manner even though the simplified model was barely different from the full model [$\chi^2(9) = 17.18$, $p = 0.046$]. However, since the full model for d' yielded a worse fit to the data [Akaike information criterion (AIC) = 999.90, Bayesian information criterion (BIC) = 1082.1] compared to the simplified model (AIC = 999.08, BIC = 1042.4), the results of the simplified model were reported here.

Table IV shows the regression coefficients for the significant effects only. Results from the NH group were not significant; they were not reported in Table IV. The performance of the CI group was found to be significantly worse than that of the NH group on all three measures: CI users' baseline accuracy score was lower than that of NH listeners by a d' of about 0.87. Moreover, CI users were, on average, 704 ms slower than NH listeners.[3] Finally, CI users, on average, accumulated information at a rate of 0.06 units/s slower than NH participants, which indicates that the increase in RTs observed for the CI group compared to the NH group was not a trade-off for increased accuracy. This means that the quality of information accrued by the CI group until they were required to give a decision was poorer compared to that of the information accumulated by the NH group.

The effect of ΔVTL was different for each group only for the d' data, as indicated by the significant interaction effect. For all other measures, all remaining effects and interactions were non-significant ($p > 0.051$).

2. NH listeners

For the NH group, no effect of $\Delta F0$, ΔVTL, or their interaction was seen on any of the three performance measures ($p > 0.20$). This indicates that the task may have been quite easy for the NH group since no further benefit on any performance measure could be drawn from the voice differences between target and masker.

3. CI listeners

For the CI group, only the VTL manipulation was found to significantly affect the d' and the drift rate ($p > 0.11$ for all other predictor variables), but not RTs ($p = 0.055$). CI users' accuracy scores dropped by an average of about 0.5 in d' per octave increase (12-st increase) in ΔVTL, and they were 0.02 units/s slower in giving a correct response for an octave increase in ΔVTL.

C. Discussion

For NH listeners, the data from experiment 1 revealed that both increasing the masker's $F0$ and shortening its VTL relative to the target speaker improved the word-by-word intelligibility of the target sentence. However, the data from experiment 2 demonstrated that overall comprehension of the target sentence as measured by the particular SVT materials chosen here, and under the specific TMR tested, did not appear to be affected by either increasing the masker's $F0$ or shortening its VTL relative to the target. Although a trend for improvement in comprehension performance as a function of increasing $\Delta F0$ or ΔVTL could be seen in the data (Fig. 4), this trend was not significant. These findings indicate that the setup for the SVT might not have been adverse enough for the NH participants, such that they mostly performed nearly at ceiling levels and hence no additional benefit could be drawn from the voice cue differences.

TABLE IV. Coefficients obtained from fitting a linear mixed-effects model to the d' and drift rate data, and a GLMM to the RT data. For conciseness, only significant effects are provided. *T*-tests reported for d' and the drift rate use Satterthwaite's approximation. *T*-values reported for RTs are obtained from the GLMM fit using maximum likelihood with Laplace approximation. Significance codes: $p < 0.05$ '*'; $p < 0.01$ '**'; $p < 0.001$ '***'.

	Fixed effect coefficient	d'	RT	Drift rate
Overall effect of group	Intercept	$\beta = 2.29$, SE = 0.18, $t(52.40) = 12.62$, $p < 0.001$***	$\beta = -0.98$, SE = 0.06, $t = -16.93$, $p < 0.001$***	$\beta = 0.12$, SE = 0.01, $t(54.80) = 11.45$, $p < 0.001$***
	group	$\beta = -0.87$, SE = 0.25, $t(52.40) = -3.36$, $p < 0.01$**	$\beta = 0.40$, SE = 0.08, $t = 4.92$, $p < 0.001$***	$\beta = -0.06$, SE = 0.02, $t(54.80) = -3.88$, $p < 0.001$***
	vtl × group	$\beta = -0.49$, SE = 0.20, $t(519.00) = -2.52$, $p = 0.012$*	—	—
CI group	Intercept	$\beta = 1.41$, SE = 0.29, $t(16.07) = 4.85$, $p < 0.001$***	$\beta = -0.57$, SE = 0.05, $t = -12.13$, $p < 0.001$***	$\beta = 0.06$, SE = 0.01, $t(16.04) = 4.61$, $p < 0.001$***
	vtl	$\beta = 0.50$, SE = 0.21, $t(17.94) = -2.36$, $p = 0.03$*	—	$\beta = -0.02$, SE = 0.01, $t(18.36) = -2.91$, $p < 0.01$**

For CI users, the data from experiment 1 revealed that both increasing the masker's $F0$ and shortening its VTL relative to those of the target speaker deteriorated the word-by-word intelligibility of the target sentence. The data from experiment 2 revealed no significant effect of $\Delta F0$ on either accuracy in d', RT, or drift rate data for the CI group. The findings of these two experiments revealed no positive benefit from $F0$ differences between two competing talkers for CI listeners, in line with the effects reported by Stickney *et al.* (2004) and Stickney *et al.* (2007), but still contradicting the findings of Pyschny *et al.* (2011). One reason for the emergence of a benefit of $\Delta F0$ in the Pyschny *et al.* study may be attributed to their high-performing bimodal CI group, as previously explained. In the current study, bimodal CI users were tested without their HA and had their HA ear blocked during testing. However, in the Pyschny *et al.* study, it is not clear whether their bimodal CI users had their HA ear blocked during testing in the CI-only condition. Thus, the discrepancy between the findings of the present study and those of Pyschny *et al.* may be attributed to the presence of usable residual hearing in the bimodal CI group tested by Pyschny *et al.*

Contrary to the effect of ΔVTL in the NH group, the effect of ΔVTL for the CI group remained consistent throughout both experiments 1 and 2: in experiment 1, shortening the masker's VTL relative to that of the target yielded systematically worse SoS intelligibility scores. This effect was persistent for SoS comprehension as measured by the SVT, in which shortening the masker's VTL led to a less accurate comprehension of the target and slower drift rates in the CI group. Hence, the remark made in Sec. III D about the increased masking effect of shorter VTLs for CI users (Fig. 3) also applies here. In addition, the same remark given in experiment 1 regarding the possible effect of TMR on the difference between the performance of the NH and CI groups also applies here.

Taken together, the results from experiments 1 and 2 revealed that CI users did not benefit from the voice differences introduced in this study between two competing talkers, such that increasing the masker's $F0$ did not lead to a positive benefit while shortening the masker's VTL yielded a decrement in performance. This means that, under the TMR conditions tested in the current study, certain voice differences that were found to be useful for NH listeners in understanding speech in the presence of background talkers were not necessarily beneficial or even slightly detrimental for CI users.

A possible explanation for this lack of benefit could be that the CI users tested in this experiment had insufficient sensitivity to $F0$ and VTL differences. This question was addressed in the following experiment.

V. EXPERIMENT 3: SENSITIVITY TO *F*0 AND VTL DIFFERENCES

A. Rationale

Experiments 1 and 2 revealed large differences between how NH and CI listeners benefit in SoS from voice differences between two concurrent speakers. NH listeners were found to benefit from both $F0$ and VTL differences between two competing talkers, while CI users were shown not to draw any benefit from such voice differences.

Because the effects reported in experiments 1 and 2 described the behavior of the CI participants as a group, it was of interest to investigate individual differences within the participants. In other words, one of the aims of this experiment was to quantify whether participants who benefited on the individual level from $F0$ and VTL differences in SoS had higher sensitivities to these two cues compared to participants who did not benefit from those voice cue differences.

The literature shows that, on the one hand, NH listeners are quite sensitive to small $F0$ and VTL differences, as was demonstrated by their low JNDs (Gaudrain and Başkent, 2018), and can utilize these two cues to categorize the gender of a speaker (Fuller *et al.*, 2014; Meister *et al.*, 2016). On the other hand, CI users are less sensitive to both $F0$ and VTL differences (Gaudrain and Başkent, 2018), and they are only able to utilize $F0$ cues (and not VTL) to categorize the gender of a speaker (Fuller *et al.*, 2014; Meister *et al.*, 2016). Because CI users, on average, have low VTL sensitivity, coupled with their inability to utilize this cue to perform gender categorization, their lack of benefit from VTL differences observed both in SoS intelligibility scores (experiment 1) and SoS comprehension performance (experiment 2) may be related to their VTL sensitivity.

Hence, this experiment measured CI users' $F0$ and VTL sensitivity using JNDs (similar to Gaudrain and Başkent, 2018) and investigated whether they were correlated with (1) the benefit in and (2) overall average SoS intelligibility (experiment 1) and comprehension performance (experiment 2). The benefit here is defined as the slopes for $\Delta F0$ and ΔVTL obtained from fitting the GLMMs in the results of the previous two experiments. This means a positive slope implies a benefit from increasing $\Delta F0$ and ΔVTL, while a negative slope indicates a decrement in performance from increasing $\Delta F0$ and ΔVTL.

B. Methods

1. Stimuli

Following the protocol defined in Gaudrain and Başkent (2015, 2018), stimuli for this experiment were taken from the NVA corpus (same as those mentioned Sec. II A). The NVA words were spoken by an adult native Dutch female speaker, with an average $F0$ of 242 Hz. Sixty-one consonant-vowel (CV) syllables with a duration between 142 ms and 200 ms were extracted from the words in the corpus, equalized in RMS, and set to a fixed duration of 200 ms using STRAIGHT (Kawahara and Irino, 2005).

A stimulus in this experiment was created by randomly selecting three different CV syllables from the list of 61 syllables, and appending them to form a triplet, with 50 ms of silence between each syllable and the next. In each trial, the same triplet of syllables was presented three times, 250 ms apart, with one of these presentations (target triplet) being different from the other two (reference triplets) in either $F0$

or VTL following an "odd-one-out" procedure [three-interval, three-alternative forced choice task (3I-3AFC)].

2. Procedure

JNDs were measured along the two principal axes (dashed grey horizontal and vertical lines in Fig. 1) relative to the reference female speaker at the origin of the [$\Delta F0$,ΔVTL] plane, yielding four voice vectors: (1) along positive (increasing) $F0$s with no change in VTL, (2) along negative (decreasing) $F0$s with no change in VTL, (3) along positive (elongating) VTLs with no change in $F0$, and (4) along negative (shortening) VTLs with no change in $F0$. Each of these conditions was repeated 3 times to yield a total of 12 runs per participant (4 voice vectors $\times$3 repetitions each). The order of the 12 runs was pseudo-randomly shuffled before presentation to each participant, and all 12 runs were conducted in a single session that lasted for about 2 h.

Each JND for a given condition was obtained using a two-down one-up adaptive procedure yielding 70.7%-correct responses on the psychometric function (Levitt, 1971). The initial trial started at a voice difference of 12 st between the reference and target triplets along one of the four voice vectors highlighted above. The voice of the two reference triplets was identical and always that of the original female speaker, and participants were asked to select the target triplet that had a different voice relative to the other two.

After each two successive correct responses, the absolute difference between the reference and target triplets decreased by a step size of 4 st. After a single incorrect response, the voice difference was increased by the same step size. If the voice difference became smaller than twice the step size, the step size was reduced by a factor of $\sqrt{2}$. The run terminated after eight reversals, and the JND was calculated as the mean voice difference, in semitones, between the target and reference triplets obtained in the last six reversals.

Before actual data collection, two training runs were provided for each participant to familiarize them with the test procedure. During training, different voices than the ones used during actual testing were selected: one voice was along a vector in the top-right quadrant of Fig. 1 ($\Delta F0 = +12$ st, ΔVTL $= -7$ st), and the other was along a different voice vector in the bottom-left quadrant of Fig. 1 ($\Delta F0 = -12$ st, ΔVTL $= +3.8$ st). Each training run was programmed to end after only six trials, irrespective of whether the adaptive procedure converged or not. Visual feedback was always provided during both training and testing, indicating to the participant whether the interval they selected was correct or not.

C. Results

1. Raw JNDs

Figure 5 shows the raw JNDs obtained for the CI group tested in this experiment. NH JND data from Gaudrain and Başkent (2015) were replotted on the same figure for comparison. A linear mixed-effects model was applied to the log-transformed JNDs, with JNDs as the predicted variable, voice vector and participant group as the predictors, and participant number as the random effect. A type III ANOVA applied to this linear model revealed that, consistent with previous findings (Gaudrain and Başkent, 2018), CI listeners had significantly higher (worse) JNDs for $F0$ [$F(1,31) = 51.47$, $p < 0.001$] and VTL [$F(1,31) = 52.62$, $p < 0.001$] compared to NH listeners. $F0$ [$F(1,31) = 0.02$, $p = 0.88$] and VTL JNDs [$F(1,31) = 0.43$, $p = 0.52$] along the positive voice vector were not significantly different than those along the negative voice vector.

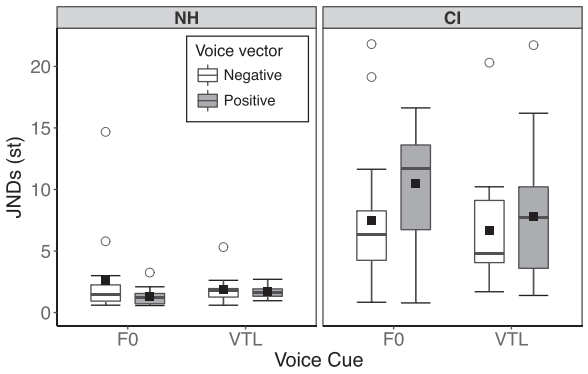

FIG. 5. $F0$ and VTL JNDs for NH (left) and CI (right) listeners. Positive (shaded bars) and negative (empty bars) voice vectors denote the principal axes of the [$\Delta F0$,ΔVTL] plane in Fig. 1, which refer to vectors from the origin along positive and negative values, respectively, of $F0$ and VTL. The boxes extend from the lower to the upper quartile, and the middle line shows the median. The whiskers show the range of the data within 1.5 times the interquartile range (IQR). The filled squares show the means, while the empty circles show the individual data outside of 1.5 times IQR.

The interaction effect between voice vector and participant group was only significant for $F0$ JNDs [$F(1,31) = 11.23$, $p = 0.002$] but not for VTL JNDs [$F(1,31) = 0.74$, $p = 0.40$]. This indicated that $F0$ JNDs along the positive voice vector (higher $F0$s) were significantly different from those along the negative voice vector (lower $F0$s) for one of the two participant groups. *Post hoc* analyses revealed that $F0$ JNDs along the positive voice vector were significantly larger (worse) than those along the negative voice vector only for the CI group [$t(17) = 3.07$, $p < 0.001$].

2. Meta-analyses

a. Relationship between JNDs and benefit from increasing $\Delta F0$ and ΔVTL. The first point of investigation in this experiment was whether the benefit from increasing $\Delta F0$ and ΔVTL from experiments 1 and 2 was correlated with the CI participants' sensitivity to $F0$ and VTL, respectively. This benefit is defined as the coefficients for $\Delta F0$ and ΔVTL from the GLMM models fitted in experiments 1 and 2, and can also be negative, in which case it would be a deficit. The benefit was plotted against the CI participants' $F0$ and VTL JNDs, as shown in Fig. 6. The top two panels show the benefit in SoS intelligibility score in Berkson per semitone increase in $\Delta F0$ (left) and ΔVTL (right) plotted against the $F0$ and VTL JNDs,

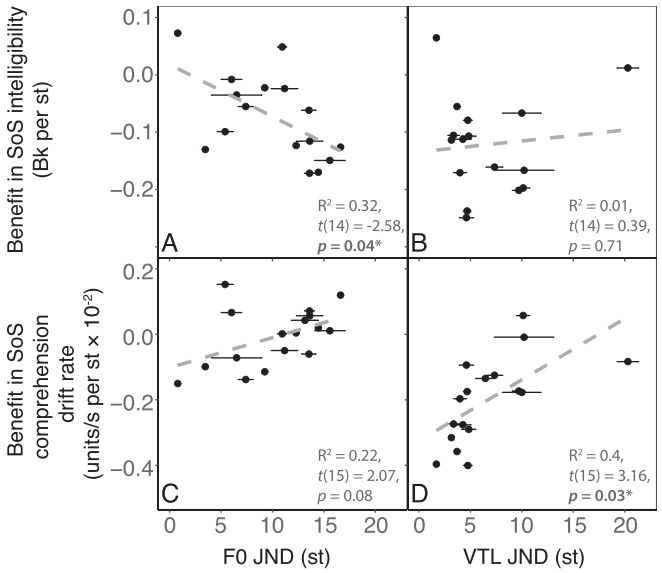

FIG. 6. Correlation between CI users' JNDs and the benefit they obtained from increasing $\Delta F0$ and ΔVTL between target and masker speakers. (A) Correlation between $F0$ JNDs and the benefit in SoS intelligibility scores [in Berkson per semitone (experiment 1)] as $\Delta F0$ increases between target and masker. (B) Correlation between VTL JNDs and the benefit in SoS intelligibility scores [in Berkson per semitone (experiment 1)] as ΔVTL increases between target and masker. (C),(D) same as (A) and (B), respectively, but for the benefit in SoS comprehension drift rate (experiment 2) measured in units per second per semitone increase. All p-values were corrected using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995). Note that negative values for the benefit denote a deficit. Error bars indicate one standard error from the mean JND.

respectively. The bottom two panels show the benefit in SoS comprehension drift rate in units/s per semitone increase in $\Delta F0$ (left) or ΔVTL (right) plotted against the $F0$ and VTL JNDs, respectively. The drift rate data were shown here because they encompass both the comprehension accuracy and RT measures, and hence serve as a more informative variable than either accuracy scores or RTs alone. A Pearson product-moment correlation coefficient was computed to assess the correlations between the benefit and the JNDs, and the p-values were corrected using the False Discovery Rate method (FDR; Benjamini and Hochberg, 1995).

b. F0 JNDs versus benefit from increasing $\Delta F0$. Only Fig. 6(A) demonstrates a negative correlation between $F0$ JNDs and the benefit in SoS intelligibility obtained by CI listeners as $\Delta F0$ increases between two simultaneous talkers. This means that, in line with what was expected, the more sensitive CI participants were to $F0$ differences (i.e., the smaller the JND), the less their SoS intelligibility scores were impaired by increasing $\Delta F0$ between target and masker speakers. In contrast, Fig. 6(C) shows a positive correlation between the benefit in drift rate obtained from increasing $\Delta F0$ and the CI participants' $F0$ JNDs. However, this correlation was non-significant, indicating that while CI participants with smaller $F0$ JNDs were less likely to experience a decrement in SoS intelligibility from increasing $\Delta F0$ differences, their $F0$ JNDs were not correlated with the benefit in SoS comprehension.

c. VTL JNDs versus benefit from increasing ΔVTL. A positive correlation was observed between VTL JNDs and the benefit from increasing ΔVTL in both experiments 1 and 2 [Figs. 6(B) and 6(D)]. Only the correlation between the benefit in drift rate and the VTL JNDs was statistically significant. This indicates that participants with larger JNDs were less likely to be affected by the masking effect introduced from increasing ΔVTL described in experiments 1 and 2.

d. Relationship between JNDs and overall performance on the SoS intelligibility and comprehension tasks. The second aim of this experiment was to assess whether $F0$ and VTL JNDs were related to the overall performance on the SoS intelligibility and comprehension tasks, rather than the relative benefit from increasing $\Delta F0$ and ΔVTL. The top plot in Fig. 7 shows the overall SoS intelligibility score for each participant (experiment 1), in percent correct (including the reference condition with no voice differences) as a function of their $F0$ and VTL JNDs. The bottom plot shows the average SoS comprehension drift rate per participant (experiment 2) also as a function of their $F0$ and VTL JNDs. The dashed lines in both plots indicate the region where typical $F0$ and VTL differences that are useful for gender categorization lie (Fuller *et al.*, 2014), i.e., a range of $F0$ up to 12 st and a range of VTL up to 3.8 st. Notice that for both data sets, participants who were sensitive to both $F0$

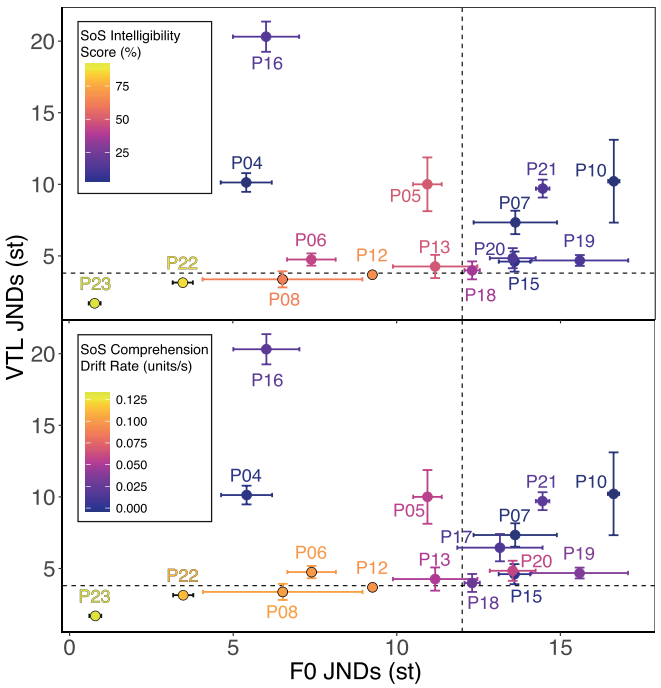

FIG. 7. (Color online) Multi-correlation plots between $F0$ JNDs, VTL JNDs, and average SoS intelligibility (top) or comprehension drift rate (bottom) per participant. The vertical dashed line indicates the range of natural $F0$ differences between typical male and female speakers (up to about 12 st). The horizontal dashed line indicates the range of natural VTL differences between typical male and female speakers (up to about 3.8 st). Horizontal and vertical error bars represent one standard error from the mean $F0$ and mean VTL JND, respectively, across the three repetitions of the JND measurement.

and VTL within this range of voice differences were the ones who tended to perform well on both the SoS intelligibility and comprehension tasks (bottom left corner). On the other hand, participants who were sensitive to either *F*0 or VTL differences, but not both, did not perform well on both tasks. A simple linear regression model with the logit of the average SoS intelligibility score as the predicted variable and *F*0 and VTL JNDs as the two predictors was significantly different than the model when either only *F*0 JNDs [$F(1,13) = 4.94$, $p = 0.045$] or only VTL JNDs [$F(1,13) = 8.46$, $p = 0.01$] was the sole predictor variable. The model with both *F*0 and VTL as predictors was able to explain 63.9% of the variance [Adjusted (Adj.) $R^2 = 0.64$ versus Adj. $R^2 = 0.41$ for *F*0 JNDs alone and Adj. $R^2 = 0.24$ for VTL JNDs]. The same type of effect was seen for the SoS comprehension drift rate, such that both *F*0 and VTL JNDs were significantly better predictors together than either only *F*0 JNDs [$F(1,14) = 10.41$, $p = 0.006$] or only VTL JNDs [$F(1,14) = 17.66$, $p < 0.001$]. This means that participants who were sensitive to both *F*0 and VTL differences (and not only one of them) were those who performed better on both the SoS intelligibility and comprehension tasks.

e. Relationship between JNDs and clinical measures. The data revealed that CI users who were more sensitive to both *F*0 and VTL differences were the ones who were, on average, more likely to perform better on both the SoS intelligibility and comprehension tasks. This raises the question of whether high performers simply perform well on all tasks rather than there being a relationship between their JNDs and SoS perception *per se*. To check for this, participants' NVA scores in quiet (see Table I) were correlated against their SoS intelligibility scores, SoS comprehension drift rate, and JNDs. The NVA scores in quiet were found to be positively correlated with the overall SoS intelligibility scores [$R^2 = 0.43$, $t(14) = 3.23$, $p = 0.006$] and SoS comprehension drift rate [$R^2 = 0.59$, $t(15) = 4.69$, $p < 0.001$], as can be expected, since participants who have poor speech intelligibility in quiet would also be more likely to suffer from the introduction of an interfering background talker. However, while the NVA scores were correlated with the VTL JNDs [$R^2 = 0.29$, $t(16) = -2.58$, $p = 0.02$], they were not correlated with the *F*0 JNDs [$R^2 = 0.08$, $t(16) = -1.17$, $p = 0.26$]. This suggests that CI users' voice cue perception is specifically related to their overall SoS intelligibility and comprehension, rather than their overall performance on other tasks, such as speech intelligibility in quiet.

The follow-up question to these findings is whether this relationship between SoS perception (both intelligibility and comprehension) and voice cue perception, could be predicted from participants' clinical measures, such as the dynamic range, the duration of CI use, the duration of hearing loss, or the age of the participants. Of these clinical measures, only the duration of hearing loss could predict the VTL JNDs [$R^2 = 0.29$, $t(16) = 2.55$, $p = 0.02$], but not the SoS intelligibility scores [$R^2 = 0.25$, $t(14) = -2.17$, $p = 0.05$], the SoS comprehension drift rate [$R^2 = 0.05$, $t(15) = -0.92$, $p = 0.37$], nor the *F*0 JNDs [$R^2 = 0.01$, $t(16) = -0.31$, $p = 0.76$]. No other clinical measures from the ones obtained in this study were found to be correlated with either the *F*0 JNDs, VTL JNDs, or SoS scores ($p > 0.13$). This indicates that it is difficult to predict which participants would have good voice cue sensitivity and SoS intelligibility and comprehension only from clinically available data.

D. Discussion

This experiment was designed to address the second and third research questions of this study: (1) whether sensitivity to *F*0 and VTL differences is related to the benefit in SoS intelligibility and comprehension as a function of increasing the difference in *F*0 and VTL between two competing talkers, and (2) whether this sensitivity was also related to each participant's overall performance on each of the intelligibility and comprehension tasks. Thus, *F*0 and VTL JNDs were measured for each CI participant and their correlations with the SoS performance measures from experiments 1 and 2 were explored.

The data revealed that *F*0 JNDs were negatively correlated with the benefit from increasing $\Delta F0$ between masker and target speakers. This means that CI users who were more sensitive to differences in *F*0 were the ones who were more likely to benefit in SoS intelligibility (experiment 1) from differences in *F*0 between two concurrent speakers. However, the benefit in SoS comprehension drift rate as $\Delta F0$ between masker and target was increased (experiment 2) was not significantly correlated with the *F*0 JNDs. These findings indicate that the presence or lack of correlations between the benefit and *F*0 JNDs may be task-related.

In contrast to the slight negative correlations observed for *F*0, a positive correlation was observed between VTL JNDs, and the benefit in both SoS intelligibility and comprehension performance as ΔVTL was increased between target and masker. This means that, counterintuitively, CI participants who were more sensitive to VTL differences were more likely to suffer from increasing ΔVTL between the two concurrent speakers. This means that being more sensitive to VTL differences may increase the sensitivity to the masking effect imposed by shortening the VTL of the masker relative to that of the target.

One possible reason for the emergence of this effect may be that participants who already started with a high baseline performance in SoS may have had no room for additional improvement with increasing ΔVTL (see Appendix B for individual data). Thus, they may have ended up experiencing a decrement in SoS performance as a function of increasing ΔVTL, as this was the only direction for their SoS scores to go to from ceiling. This, in fact, appeared to be the case only for SoS comprehension, but not intelligibility, when the baseline performance for the SoS intelligibility and comprehension tasks (as estimated by the intercept of the linear model) were investigated for correlations with the benefit from increasing ΔVTL [for intelligibility: $R^2 = 0.24$, $t(14) = 1.66$, $p = 0.12$; for comprehension: $R^2 = 0.87$, $t(15) = -10.04$, $p < 0.0001$]. Repeating the GLMM analyses with only the participants

who were not at floor or ceiling revealed similar effects as those reported in the results section for the group average. These findings indicate that the effects reported in this study are not largely dictated by floor or ceiling effects.

The relationship between the average SoS performance (across all *F*0 and VTL differences) and JNDs was much clearer: CI users who were more sensitive to both *F*0 and VTL differences were, on average, more likely to perform better on both the SoS intelligibility and comprehension tasks. This relationship was found to be particular to the SoS and JND tasks, and was not merely a result of having participants who performed well irrespective of task administered, since the NVA scores in quiet could not predict performance on all measures of voice cue sensitivity and SoS performance. In addition, the relationship between voice cue sensitivity and SoS performance could not be predicted from available clinical data, meaning that it is challenging to predict participants' sensitivity to voice cues by only looking at the clinical data during recruitment. Thus, selecting a wide range of NVA scores during the recruitment phase of this study allowed for observing this relationship between JNDs and SoS performance.

These findings, however, cannot be generalized to all types of voice differences, since, in this study, only a specific type of voice manipulation was applied. All *F*0 and VTL changes used here encompassed only child-like voices (top-right quadrant) from the whole [$\Delta F0$,ΔVTL] plane of possible values. Thus, it is unknown whether a similar pattern of results would be seen if *F*0 and VTL were manipulated to sound more male-like compared to the reference (i.e., fall in the lower-left quadrant of the [$\Delta F0$,ΔVTL] plane). In fact, it is expected that if the masker's VTL was elongated relative to that of the target speaker, CI users should obtain a benefit in SoS performance. This is because elongating the masker's VTL is expected to result in a compression of the masker's spectral envelope toward lower frequencies. Thus, it is expected that this type of voice manipulation may yield a higher TMR across electrodes (less interaction between target and masker stimulation patterns), which is opposite to the effect shown in Fig. 3.

Additionally, it is unknown if a similar pattern of results would be observed at different TMRs. Previous work has demonstrated that the size of the benefit from voice differences in SoS scenarios likely depends on the TMR (e.g., Darwin *et al.*, 2003; Stickney *et al.*, 2004). Thus, it would be beneficial to investigate whether the effects observed in the current study would persist at different TMRs.

It is worth commenting on the relatively large difference between the average ages of the NH and CI groups tested in this study. The age difference was caused by a number of factors. First, the young NH listeners were recruited only as a control group for the methodological validation, i.e., to rule out that the detrimental effects observed in the CI group were not a result of the voice manipulations *per se*. Second, the recruitment of younger CI users was not sufficiently practical within the time frame of this study because of the overlap of potential testing times and their work schedules. Nevertheless, the ages of the CI participant sample recruited spanned a large range (from 33.3 yr to 76.1 yr), which allowed us to investigate whether the effect of age within this sample was a potential confound to the results obtained.

Regarding the JND data, there are contradictory findings in the literature regarding the effect of age on *F*0 differences. For example, Souza *et al.* (2011) reported evidence that younger NH listeners were more sensitive to *F*0 differences under noise-vocoded conditions compared to older NH listeners. Contrary to this, Gaudrain and Başkent (2015) investigated the effects of vocoding on *F*0 and VTL JNDs in NH listeners, and also assessed whether the age range of their NH participant group influenced the pattern of results observed. While it was not a systematic study of the effect of age on JNDs, the authors reported that the large age range of their NH group (19–63 yr) did not significantly affect the JNDs measured in that study. Building on that idea because JNDs in the current study were measured only for the CI group and not the NH group, a linear regression model was fitted to the log-transformed *F*0 and VTL JNDs with age as a fixed-effect predictor. These analyses revealed that age was not a significant predictor of either *F*0 [$\beta = 0.02$, SE $= 0.01$, $F(1,16) = 1.07$, $p = 0.31$] or VTL JNDs [$\beta = 0.003$, SE $= 0.01$, $F(1,16) = 0.08$, $p = 0.78$]. Based on these findings, it seems unlikely that age was the dominating factor contributing to the pattern of results observed, at least not in a systematic manner. However, a more systematic study needs to be performed before this potential effect of age on JNDs can be comprehensively identified or ruled out.

Regarding the SoS intelligibility data, previous studies have demonstrated that age can impact speech intelligibility under adverse listening conditions. For example, older NH participants were shown to have lower speech intelligibility in the presence of background noise (e.g., Gordon-Salant and Fitzgibbons, 1999) or competing talkers (e.g., Başkent *et al.*, 2014; Bergman *et al.*, 1976; Tun *et al.*, 2002) compared to younger NH participants. Since aging effects have been reported in the literature for NH listeners, these effects are expected to be even more highlighted when comparing the performance of young NH listeners to that of older CI users. Directly supporting this idea, Bhargava *et al.* (2016) have shown that substantial differences in the intelligibility of interrupted speech existed between older CI participants and younger NH listeners listening to vocoder simulations. When the authors tested a second NH sample with participants who were age-matched to the CI group, the age-matched NH group's performance under vocoded conditions approached that of the CI group. In the current study, whether the effect of age could have confounded the benefit in SoS intelligibility results was investigated both within the CI group itself and, since NH listeners were also tested with SoS, also between the CI and NH groups. Within the CI group, a logistic regression model was fitted to the binary per-word score using the full factorial model, as shown in Eq. (2), with the added effect of participant age. The full

factorial model accounted for the interaction between the effects of $F0$ and VTL with age on SoS intelligibility. If the interaction with age is significant, this would mean that the effects of $F0$ and VTL on SoS intelligibility scores in CI users would change depending on age. However, the logistic regression model revealed no significant effect of age as a fixed-factor [$\beta = -1.51$, SE = 1.76, $z = -0.86$, $p = 0.39$] nor significant interactions between any of the fixed effects and age ($p > 0.15$).

Investigating the effect of age between the CI and NH groups, a logistic regression model was fit to the entire SoS intelligibility dataset, with the full factorial specification [as provided in Eq. (1)] of the effects of $F0$, VTL, age, and participant group. The logistic regression revealed no effect of age as a fixed-factor [$\beta = -4.33$, SE = 13.20, $z = -0.33$, $p = 0.74$], and no significant interactions involving age ($p > 0.19$). Similar to the analyses performed on the JND data, these analyses revealed that age did not appear to significantly modulate the difference between the NH and CI results observed. Nevertheless, since the literature reports that age could have contributed to the difference between the effects for NH and CI listeners reported in this study, the effect of age should be explicitly investigated in a follow-up study in a systematic manner, by including age-matched NH controls or younger CI participants.

It is important to note that the effects of voice cues on SoS perception observed for the CI group, although statistically significant, were in fact small. This may be due to the considerable inter-subject variability in the performance of the CI group tested (see individual data in Figs. 8 and 9) compared to that of the NH listeners recruited for this study (e.g., z-statistic for SoS intelligibility as a function of $\Delta F0$ is 8.86 for NH versus −3.00 for CI users, and as a function of ΔVTL is 11.56 for NH listeners versus −4.5 for CI users). Thus, whether more substantial effects may be observed for a larger, more homogenous CI group (i.e., whose performance is away from floor and ceiling) remains currently unknown.

Since sensitivity to both $F0$ and VTL cues was found to be related to overall SoS performance, the question then arises of whether improving one measure would necessarily lead to an improvement in the other. In a previous study, VTL JNDs were found to depend on the frequency-to-electrode allocation mapping in vocoder simulations of CI processing (El Boghdady *et al.*, 2018). It remains to be seen whether implant parameters, such as the frequency-to-electrode allocation mapping or the coding strategy, could help improve SoS performance in addition to JNDs.

VI. CONCLUSION

This study was designed to address three research questions: (1) Do CI users benefit in SoS scenarios from $F0$ and VTL differences between two competing talkers in a manner similar to NH listeners? (2) Is this benefit related to their sensitivity to $F0$ and VTL differences? (3) Could their overall average SoS performance be related to their $F0$ and VTL sensitivity? The results from this study revealed that: (1) Contrary to NH listeners, CI listeners do not benefit from $F0$ differences between two concurrent speakers, while they experience a decrement in performance as the masker's VTL is shortened relative to that of the target. (2) The effect on SoS perception from increasing the relative difference in $F0$ and VTL between two competing talkers is related to the CI users' sensitivity to these two voice cues. (3) CI users' overall average performance on a variety of SoS tasks can be mainly predicted by their sensitivity to both $F0$ and VTL differences. These findings indicate that $F0$ and VTL JNDs may serve as useful methods to investigate the effectiveness of new speech coding strategies since they are directly related to SoS performance.

ACKNOWLEDGMENTS

The work presented here was jointly funded by Advanced Bionics (AB), the University Medical Center Groningen (UMCG), and the PPP-subsidy of the Top Consortia for Knowledge and Innovation of the Ministry of Economic Affairs. The study was additionally supported by a Rosalind Franklin Fellowship from the University Medical Center Groningen, University of Groningen, and the VICI Grant No. 016.VICI.170.111 from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw). This work was conducted in the framework of the LabEx CeLyA ("Centre Lyonnais d'Acoustique," ANR-10-LABX-0060/ANR-11-IDEX-0007) operated by the French National Research Agency, and is also part of the research program of the Otorhinolaryngology Department of the University Medical Center Groningen: Healthy Aging and Communication. The authors would like to especially thank Bert Maat, Emile de Klein, Rolien Free, and Gerda Boven for their help with recruiting CI participants, contacting external clinics, and obtaining the clinical measures reported in the CI users' demographics table; Keel-, Neus-, en Oorheelkunde (KNO) clinics from Leiden UMC, UMC Maastricht, Radboud UMC, and UMC Utrecht for the help with recruiting CI participants; Enja Jung for recording the Dutch SVT corpus from a female speaker and processing the stimuli; Marieke van Vugt for her advice on RT paradigms and how to use drift-diffusion models; Dutch student assistants Julia Verbist, Charlotte de Blecourt, Fergio Sismono, and Britt Bosma for their help with conducting the experiments and online scoring the verbal responses from the CI users; Paolo Toffanin for his help with stimuli calibration; Jeanne Clarke for her help with audiometric measurements; all colleagues who helped pilot this study, all the NH and CI participants who volunteered, and all the staff of the KNO clinic at the University Medical Center Groningen (UMCG).

APPENDIX A: TRAINING SENTENCES DEVELOPED FOR THE SVT

The true/false sentence pairs introduced to the SVT material to be used for training purposes are shown in Table V.

TABLE V. True/false sentences added to the SVT material and used only to train participants to the nature of the task.

True		False	
Dutch	English translation	Dutch	English translation
Muizen zijn klein	Mice are small	Gras is meestal rood	Grass is mainly red
Auto's zijn meestal sneller dan fietsers	Cars are usually faster than bikes	Konijnen eten olifanten op	Rabbits eat elephants
Zebra's hebben zwarte en witte strepen	Zebras have black and white stripes	De zon is koud	The Sun is cold
Leraren staan voor de las	Teachers stand at the front of the class	Duitsland is een land op de maan	Germany is a country on the moon

APPENDIX B: INDIVIDUAL DATA

Individual data for each participant are shown for the SoS intelligibility and comprehension tasks in Figs. 8 and 9, respectively.

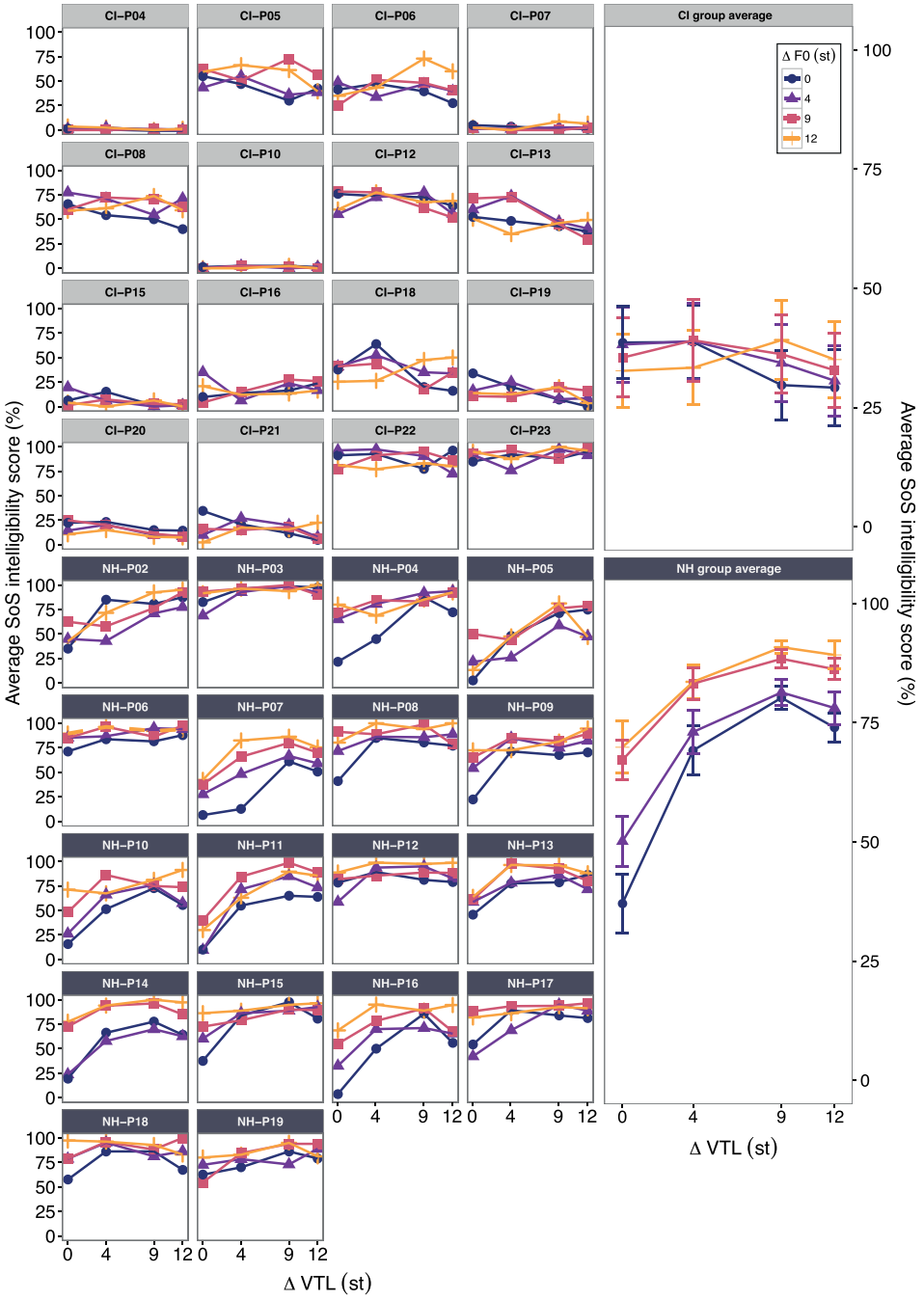

FIG. 8. (Color online) Individual (small panels) and group (large panels) SoS intelligibility scores (experiment 1) plotted against ΔVTL for each value of Δ*F*0. (Top) (light labels) CI data. (Bottom) (dark labels) NH data.

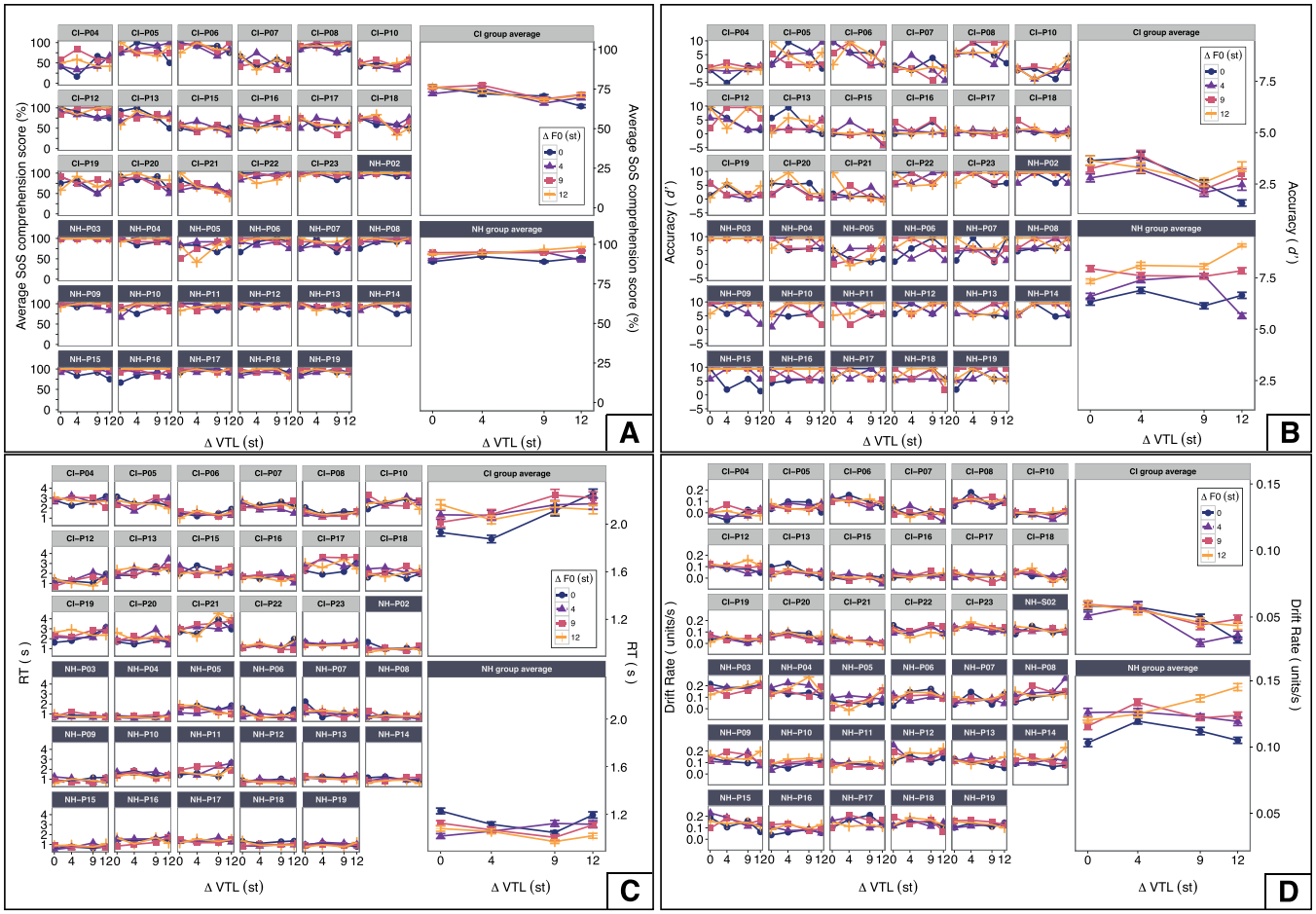

FIG. 9. (Color online) CI (light grey panels) and NH (dark grey panels) performance on the SoS comprehension task (SVT; experiment 2). Individual data are shown in the small panels on the left labeled by each participant's number, while group data are shown in the larger panels on the right. (A) Average percent correct scores as a function of increasing Δ*F*0 and ΔVTL. (B) Accuracy scores in *d*′. (C) RT data in seconds. (D) Drift rate data in arbitrary units per second.

[1]The term "gender," as used in the context of this study, denotes the classical categorization of a speaker's voice as belonging to either a cisgender male or cisgender female [a person whose perceived gender identity corresponds to their assigned sex at birth (American Psychological Association, 2015)].

[2]Berkson (Bk) is a dimensionless unit named after Joseph Berkson (1899–1982) who popularised the use of log odds-ratios, where the odds-ratio is the ratio of correct to incorrect responses in logistic regression. The Berkson unit, defined as $\log_2$(odds-ratio), serves to linearize the logistic scale such that a constant change along the Bk scale corresponds to a constant change on the decibel scale (see, for example, Hilkhuysen *et al.*, 2012, for a description). An increase by 1 Bk unit is equivalent to a doubling of the number of correct responses when the number of incorrect responses is fixed, while an increase in the raw log(odds-ratio) by 1 results in an increase in the odds-ratio by a factor of 2.7183, which is less intuitive. Thus, to convert the log(odds-ratio) to units of Bk [$\log_2$(odds-ratio)], the log(odds-ratio) needs to be divided by log(2). The benefit in Bk reported here was calculated by converting the normalized coefficients for each variable back into units of semitones and dividing that quantity by log(2).

[3]This difference in baseline performance between the two participant groups was computed by substituting the linear regression coefficients (Table IV) into the linear regression model for RTs.

Abercrombie, D. (**1967**). *Elements of General Phonetics* (Edinburgh University Press, Edinburgh), Vol. 203.

Adank, P., and Janse, E. (**2009**). "Perceptual learning of time-compressed and natural fast speech," J. Acoust. Soc. Am. **126**, 2649–2659.

American Psychological Association (**2015**). "Guidelines for psychological practice with transgender and gender nonconforming people," Am. Psychol. **70**, 832–864.

Assmann, P. F., and Summerfield, Q. (**1990**). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. **88**, 680–697.

Baddeley, A. D., Emslie, H., and Nimmo-Smith, I. (**1992**). *The Speed and Capacity of Language-Processing Test* (Thames Valley Test Company, Bury St Edmunds, UK).

Baddeley, A., Gardner, J. M., and Grantham-McGregor, S. (**1995**). "Cross-cultural cognition: Developing tests for developing countries," Appl. Cognit. Psychol. **9**, S173–S195.

Baer, T., Moore, B. C. J., and Gatehouse, S. (**1993**). "Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: Effects on intelligibility, quality, and response times," J. Rehabil. Res. Dev. **30**, 49–72.

Başkent, D., and Gaudrain, E. (**2016**). "Musician advantage for speech-on-speech perception," J. Acoust. Soc. Am. **139**, EL51–EL56.

Başkent, D., Gaudrain, E., Tamati, T. N., and Wagner, A. (**2016**). "Perception and psychoacoustics of speech in cochlear implant users," in *Scientific Foundations of Audiology: Perspectives from Physics, Biology, Modeling, and Medicine*, edited by A. T. Cacace, E. de Kleine, A. Genene-Holt, and P. van Dijk (Plural Publishing, Inc, San Diego, CA), pp. 285–319.

Başkent, D., Luckmann, A., Ceha, J., Gaudrain, E., and Tamati, T. N. (**2018**). "The discrimination of voice cues in simulations of bimodal electro-acoustic cochlear-implant hearing," J. Acoust. Soc. Am. **143**, EL292–EL297.

Başkent, D., van Engelshoven, S., and Galvin, J. J. III (**2014**). "Susceptibility to interference by music and speech maskers in middle-aged adults," J. Acoust. Soc. Am. **135**, EL147–EL153.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Software **67**, 1–48.

Benjamini, Y., and Hochberg, Y. (**1995**). "Controlling the false discovery rate: A practical and powerful approach to multiple testing," J. R. Stat. Soc. Ser. **57**, 289–300.

Bergman, M., Blumenfeld, V. G., Cascardo, D., Dash, B., Levitt, H., and Margulies, M. K. (**1976**). "Age-related decrement in hearing for speech: Sampling and longitudinal studies," J. Gerontol. **31**, 533–538.
Best, V., Keidser, G., Buchholz, J. M., and Freeston, K. (**2016**). "Development and preliminary evaluation of a new test of ongoing speech comprehension," Int. J. Audiol. **55**, 45–52.
Bhargava, P., Gaudrain, E., and Başkent, D. (**2016**). "The intelligibility of interrupted speech: Cochlear implant users and normal hearing listeners," J. Assoc. Res. Otolaryngol. **17**, 475–491.
Bosman, A. J., and Smoorenburg, G. F. (**1995**). "Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment," Audiology **34**, 260–284.
Brokx, J., and Nooteboom, S. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonetics **10**, 23–36.
Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2009**). "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers," J. Acoust. Soc. Am. **125**, 4006–4022.
Carlyon, R. P., and Shackleton, T. M. (**1994**). "Comparing the fundamental frequencies of resolved and unresolved harmonics: Evidence for two pitch mechanisms?," J. Acoust. Soc. Am. **95**, 3541–3554.
Chiba, T., and Kajiyama, M. (**1941**). *The Vowel: Its Nature and Structure* (Tokyo Kaiseikan, Tokyo).
Clarke, J., Gaudrain, E., Chatterjee, M., and Başkent, D. (**2014**). "T'ain't the way you say it, it's what you say–Perceptual continuity of voice and top–down restoration of speech," Hear. Res. **315**, 80–87.
Cullington, H. E., and Zeng, F.-G. (**2008**). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," J. Acoust. Soc. Am. **123**, 450–461.
Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.
Drullman, R., and Bronkhorst, A. W. (**2004**). "Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers," J. Acoust. Soc. Am. **116**, 3090–3098.
El Boghdady, N., Başkent, D., and Gaudrain, E. (**2018**). "Effect of frequency mismatch and band partitioning on vocal tract length perception in vocoder simulations of cochlear implant processing," J. Acoust. Soc. Am. **143**, 3505–3519.
Fant, G. (**1960**). *Acoustic Theory of Speech Perception* (Mouton, The Hague).
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522.
Fuller, C. D., Gaudrain, E., Clarke, J. N., Galvin, J. J., Fu, Q.-J., Free, R. H., and Başkent, D. (**2014**). "Gender categorization is abnormal in cochlear implant users," J. Assoc. Res. Otolaryngol **15**, 1037–1048.
Gatehouse, S., and Gordon, J. (**1990**). "Response times to speech stimuli as measures of benefit from amplification," Br. J. Audiol. **24**, 63–68.
Gaudrain, E., and Başkent, D. (**2015**). "Factors limiting vocal-tract length discrimination in cochlear implant simulations," J. Acoust. Soc. Am. **137**, 1298–1308.
Gaudrain, E., and Başkent, D. (**2018**). "Discrimination of voice pitch and vocal-tract length in cochlear implant users," Ear Hear. **39**, 226–237.
Gordon-Salant, S., and Fitzgibbons, P. J. (**1999**). "Profile of auditory temporal processing in older listeners," J. Speech. Lang. Hear. Res. **42**, 300–311.
Green, D., and Swets, J. (**1966**). *Signal Detection Theory and Psychophysics* (Wiley, New York).
Hecker, M. H., Stevens, K. N., and Williams, C. E. (**1966**). "Measurements of reaction time in intelligibility tests," J. Acoust. Soc. Am. **39**, 1188–1189.
Hilkhuysen, G., Gaubitch, N., Brookes, M., and Huckvale, M. (**2012**). "Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios," J. Acoust. Soc. Am. **131**, 531–539.
Hillenbrand, J. M., and Clark, M. J. (**2009**). "The role of *f*0 and formant frequencies in distinguishing the voices of men and women," Atten., Percept. Psychophys. **71**, 1150–1166.
Ives, D. T., Smith, D. R. R., and Patterson, R. D. (**2005**). "Discrimination of speaker size from syllable phrases," J. Acoust. Soc. Am. **118**, 3816–3822.
Kawahara, H., and Irino, T. (**2005**). "Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Springer, Boston, MA), pp. 167–180.
Kiessling, J., Pichora-Fuller, M. K., Gatehouse, S., Stephens, D., Arlinger, S., Chisolm, T., Davis, A. C., Erber, N. P., Hickson, L., Holmes, A., Rosenhall, U., and von Wedel, H. (**2003**). "Candidature for and delivery of audiological services: Special needs of older people," Int. J. Audiol. **42**, 92–101.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Licklider, J. (**1954**). " 'Periodicity' pitch and 'place' pitch," J. Acoust. Soc. Am. **26**, 945–945.
Lieberman, P., and Blumstein, S. E. (**1988**). "Source-filter theory of speech production," in *Speech Physiology, Speech Perception, and Acoustic Phonetics* (Cambridge University Press, Cambridge), pp. 34–50.
Lo, S., and Andrews, S. (**2015**). "To transform or not to transform: Using generalized linear mixed models to analyse reaction time data," Front. Psychol. **6**, 1171.
May, J., Alcock, K. J., Robinson, L., and Mwita, C. (**2001**). "A computerized test of speed of language comprehension unconfounded by literacy," Appl. Cognit. Psychol. **15**, 433–443.
Meister, H., Fürsen, K., Streicher, B., Lang-Roth, R., and Walger, M. (**2016**). "The use of voice cues for speaker gender recognition in cochlear implant recipients," J. Speech. Lang. Hear. Res. **59**, 546–556.
Müller, J. (**1848**). *The Physiology of the Senses, Voice, and Muscular Motion, with the Mental Faculties* (Taylor, Walton and Maberly, London).
Oxenham, A. J. (**2008**). "Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants," Trends Amplif. **12**, 316–331.
Pachella, R. G. (**1974**). "The interpretation of reaction time in human information processing research," in *Human Information Processing: Tutorials in Performance and Cognition*, edited by B. H. Kantowitz (Erlbaum Associates, Hillsdale, NJ).
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Pisoni, D. B., Manous, L. M., and Dedina, M. J. (**1987**). "Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility," Comput. Speech Lang. **2**, 303–320.
Plomp, R., and Mimpen, A. M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **18**, 43–52.
Pyschny, V., Landwehr, M., Hahn, M., Walger, M., von Wedel, H., and Meister, H. (**2011**). "Bimodal hearing and speech perception with a competing talker," J. Speech. Lang. Hear. Res. **54**, 1400–1415.
Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.
Qin, M. K., and Oxenham, A. J. (**2005**). "Effects of envelope-vocoder processing on *F*0 discrimination and concurrent-vowel identification," Ear Hear. **26**, 451–460.
R Core Team (**2017**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, available at https://www.R-project.org/ (Last viewed July 2018).
Rana, B., Buchholz, J. M., Morgan, C., Sharma, M., Weller, T., Konganda, S. A., Shirai, K., and Kawano, A. (**2017**). "Bilateral versus unilateral cochlear implantation in adult listeners: Speech-on-speech masking and multitalker localization," Trends Hear. **21**, 1–15.
Ratcliff, R. (**1978**). "A theory of memory retrieval," Psychol. Rev. **85**, 59–108.
Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (**2016**). "Diffusion decision model: Current issues and history," Trends Cognit. Sci. **20**, 260–281.
Saxton, J. A., Ratcliff, G., Dodge, H., Pandav, R., Baddeley, A., and Ganguli, M. (**2001**). "Speed and capacity of language processing test: Normative data from an older American community-dwelling sample," Appl. Neuropsychol. **8**, 193–203.
Schönbeck, Y. (**2010**). "Growth chart Dutch girls 1-21 years," TNO, Leiden, available at https://www.tno.nl/en/focus-areas/healthy-living/roadmaps/youth/pdf-growth-charts/ (Last viewed July 2018).
Schouten, J., and Bekker, J. (**1967**). "Reaction time and accuracy," Acta Psychol. **27**, 143–153.
Skuk, V. G., and Schweinberger, S. R. (**2014**). "Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender," J. Speech. Lang. Hear. Res. **57**, 285–296.

Smith, D. R. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age," J. Acoust. Soc. Am. **118**, 3177–3186.
Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**, 305–318.
Smith, D. R. R., Walters, T. C., and Patterson, R. D. (**2007**). "Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled," J. Acoust. Soc. Am. **122**, 3628–3639.
Souza, P., Arehart, K., Miller, C. W., and Muralimanohar, R. K. (**2011**). "Effects of age on *F*0-discrimination and intonation perception in simulated electric and electro-acoustic hearing," Ear Hear. **32**, 75–83.
Stevens, K. N., and House, A. S. (**1955**). "Development of a quantitative description of vowel articulation," J. Acoust. Soc. Am. **27**, 484–493.
Stickney, G. S., Assmann, P. F., Chang, J., and Zeng, F.-G. (**2007**). "Effects of cochlear implant processing and fundamental frequency on the intelligibility of competing sentences," J. Acoust. Soc. Am. **122**, 1069–1078.
Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (**2004**). "Cochlear implant speech recognition with speech maskers," J. Acoust. Soc. Am. **116**, 1081–1091.
Swanson, B., and Mauch, H. (**2006**). "Nucleus MATLAB Toolbox 420 software user manual," Cochlear Ltd., Lane Cove NSW, Australia.
Tun, P. A., O'Kane, G., and Wingfield, A. (**2002**). "Distraction by competing speech in young and older adult listeners," Psychol. Aging **17**, 453–467.
Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (**2000**). "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**, 1671–1684.
Vestergaard, M. D., Fyson, N. R., and Patterson, R. D. (**2009**). "The interaction of vocal characteristics and audibility in the recognition of concurrent syllables," J. Acoust. Soc. Am. **125**, 1114–1124.
Wagenmakers, E.-J., Van Der Maas, H. L., and Grasman, R. P. (**2007**). "An EZ-diffusion model for response time and accuracy," Psychon. Bull. Rev. **14**, 3–22.
Wickelgren, W. A. (**1977**). "Speed-accuracy tradeoff and information processing dynamics," Acta Psychol. **41**, 67–85.
Zaltz, Y., Goldsworthy, R. L., Kishon-Rabin, L., and Eisenberg, L. S. (**2018**). "Voice discrimination by adults with cochlear implants: The Benefits of early implantation for vocal-tract length perception," J. Assoc. Res. Otolaryngol. **19**, 193–209.